**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆP THỰC PHẨM TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN MÔN HỌC**

**PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO**

**ĐỀ TÀI: PHÂN TÍCH TẬP DỮ LIỆU**

*Giáo viên hướng dẫn: Đinh Nguyễn Trọng Nghĩa*

*Sinh viên thực hiện: 2001190172 - Đặng Kim Ngân*

*2001190129 - Nguyễn An Khương*

**LỜI MỞ ĐẦU**

**MỤC LỤC**

Contents

[**PHÂN CÔNG** 5](#_Toc109771062)

[**DANH MỤC HÌNH ẢNH** 5](#_Toc109771063)

[**CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN TẬP DỮ LIỆU** 8](#_Toc109771064)

[**CHƯƠNG 2: QUẢN LÝ DỮ LIỆU** 9](#_Toc109771065)

[1) Thông tin tập dữ liệu 9](#_Toc109771066)

[2) Loại bỏ dữ liệu thiếu 9](#_Toc109771067)

[3) Mã hóa dữ liệu 12](#_Toc109771068)

[4) Nhóm các giá trị và tạo biến phụ 13](#_Toc109771069)

[**CHƯƠNG 3: KIỂM ĐỊNH GIẢ THUYẾT** 15](#_Toc109771070)

[1) Phân tích phương sai ANOVA 15](#_Toc109771071)

[a) Phân tích ANOVA cho biến giải thích có 2 loại 15](#_Toc109771072)

[b) Phân tích ANOVA cho biến giải thích nhiều hơn 2 loại. 17](#_Toc109771073)

[c) Phân tích sâu ANOVA 19](#_Toc109771074)

[2) Phân tích Chi – square 20](#_Toc109771075)

[3) Tương quan Pearson 22](#_Toc109771076)

[**CHƯƠNG 4: PHÂN TÍCH HỒI QUY** 24](#_Toc109771077)

[1) Hồi quy tuyến tính giữa 2 biến định lượng 24](#_Toc109771078)

[2) Hồi quy tuyến tính giữa biến định lượng và phân loại 25](#_Toc109771079)

[3) Hồi quy đa biến 27](#_Toc109771080)

[5) Hồi quy Logistic 33](#_Toc109771081)

[a) Hồi quy Logistic đơn biến 33](#_Toc109771082)

[b) Hồi quy Logistic đa biến 35](#_Toc109771083)

[**Chương 5: Cây quyết định** 36](#_Toc109771084)

# **PHÂN CÔNG**

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ và tên** | **Công việc** |
| 2001190172 | Đặng Kim Ngân |  |
| 2001190129 | Nguyễn An Khương |  |
|  | Dương Bảo Phước |  |

# **DANH MỤC HÌNH ẢNH**

[Hình 1: Thông tin dữ liệu 10](#_Toc86773808)

[Hình 2: Tỉ lệ dữ liệu thiếu của biến 11](#_Toc86773809)

[Hình 3: Trực quan hóa dữ liệu thiếu bằng biểu đồ nhiệt 11](#_Toc86773810)

[Hình 4: In ra các dòng dữ liệu thiếu 12](#_Toc86773811)

[Hình 5: Số dòng dữ liệu thiếu 12](#_Toc86773812)

[Hình 6: Import lớp SimpleImputer từ thư viện sklearn.impute 12](#_Toc86773813)

[Hình 7: Tạo giá trị mean 13](#_Toc86773814)

[Hình 8: Thông tin dữ liệu sau khi cập nhật 13](#_Toc86773815)

[Hình 9: Mã hóa dữ liệu 13](#_Toc86773816)

[Hình 10: Kết quả sau khi mã hóa 14](#_Toc86773817)

[Hình 11: Tạo biến phụ cho Ratings 14](#_Toc86773818)

[Hình 12: Tạo biến phụ cho Year 15](#_Toc86773819)

[Hình 13: Phân tích ANOVA cho biến giải thích có 2 loại 17](#_Toc86773820)

[Hình 14: Tính trung bình và độ lệch chuẩn 17](#_Toc86773821)

[Hình 15: Phân tích ANOVA cho biến giải thích nhiều hơn 2 loại 19](#_Toc86773822)

[Hình 16: Tính trung bình và độ lệch chuẩn 19](#_Toc86773823)

[Hình 17: Hàm tukeyhsd() 20](#_Toc86773824)

[Hình 18: Kết quả phân tích sâu ANOVA 20](#_Toc86773825)

[Hình 19: Dùng hàm pandas.crosstab 21](#_Toc86773826)

[Hình 20: Tính phần trăm cột 22](#_Toc86773827)

[Hình 21: Tính Chi - square 22](#_Toc86773828)

[Hình 22: Đồ thị phân tán 23](#_Toc86773829)

[Hình 23: Tính hệ số tương quan 24](#_Toc86773830)

[Hình 24: Biểu đồ phân tán 2 biến định lượng 25](#_Toc86773831)

[Hình 25: Mô hình hồi quy tuyến tính 2 biến định lượng 26](#_Toc86773832)

[Hình 26: Mô hình hồi quy tuyến tính giữa biến phân loại và định lượng 27](#_Toc86773833)

[Hình 27: Biểu đồ cho cái nhìn trực quan cho mô hình 27](#_Toc86773834)

[Hình 28: Canh chuẩn cho Screens 28](#_Toc86773835)

[Hình 29: Canh chuẩn cho Likes 29](#_Toc86773836)

[Hình 30: Canh chuẩn cho Dislikes 29](#_Toc86773837)

[Hình 31: Mô hình hồi quy đa biến 30](#_Toc86773838)

[Hình 32: Đồ thị phân tán biểu diễn 2 biến 31](#_Toc86773839)

[Hình 33: Canh chuẩn dữ liệu cho Ratings 31](#_Toc86773840)

[Hình 34: Mô hình hồi quy bậc 1 32](#_Toc86773841)

[Hình 35: Mô hình hồi quy bậc 2 33](#_Toc86773842)

[Hình 36: Mô hình hồi quy Logistic đơn biến 34](#_Toc86773843)

[Hình 37: Tính ODD 35](#_Toc86773844)

[Hình 38: Lấy khoảng tin cậy 35](#_Toc86773845)

[Hình 39: Mô hình hồi quy Logistic đa biến 36](#_Toc86773846)

# **CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN TẬP DỮ LIỆU**

Trong tập dữ liệu “Conventional and Social Media Movies”, dữ liệu được thu thập từ các bộ phim điện ảnh năm 2014 và 2015. Nguồn dữ liệu ban đầu là từ Youtube, Twitter và IMDB (Ahmed M, Jahangir M, Afzal H, Majeed A, Siddiqi I). Và mỗi bản khảo sát đặt ra 14 câu hỏi cho một mẫu chỉ định trước (bộ phim được sản xuất trong năm 2014 và 2015). Tập dữ liệu bao gồm 231 khảo sát với các biến thể hiện giá trị của một bộ phim trong ngành điện ảnh thế giới. Có tổng cộng 14 thuộc tính trong tập dữ liệu với một số dữ liệu bị thiếu. Kích thước tập dữ liệu gồm 231 dòng và 14 cột với các thuộc tính như:

+ Movie: Tên phim

+ Year: Năm sản xuất (2014, 2015)

+ Ratings: Tỉ lệ đánh giá

+ Genre: Thể loại. (1. Action, 2. Science fiction, 3. Drama,   
4. Mystery thriller; 6. Erotic thriller; 7. Drama, mystery & thriller;   
8. Comedy; 9. Historical drama; 10. Neo-noir thriller; 12. Computer animated; 15. Horror)

+ Gross: Tổng doanh thu

+ Budget: Chi phí sản xuất

+ Screens: Số lượng rạp chiếu

+ Sequel: Phần tiếp theo

+ Sentiment: Điểm tình cảm

+ Views: Số lượt xem

+ Likes: Số lượng yêu thích

+ Dislikes: Số lượng không thích

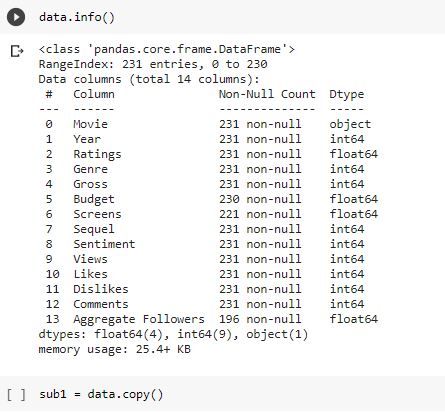
+ Comments: Số lượng nhận xét

+ Aggregate Followers: Tổng lượng người theo dõi diễn viên

Dữ liệu phân loại bao gồm tên và thể loại phim trong khi năm (Year) thuộc về dữ liệu thứ tự. Có tổng cộng 11 thuộc tính là dữ liệu số. Chỉ có hai biến duy nhất cho năm là 2014 và 2015. Có 231 biến duy nhất cho tên phim và 15 biến duy nhất đại diện cho các thể loại phim khác nhau.

# **CHƯƠNG 2: QUẢN LÝ DỮ LIỆU**

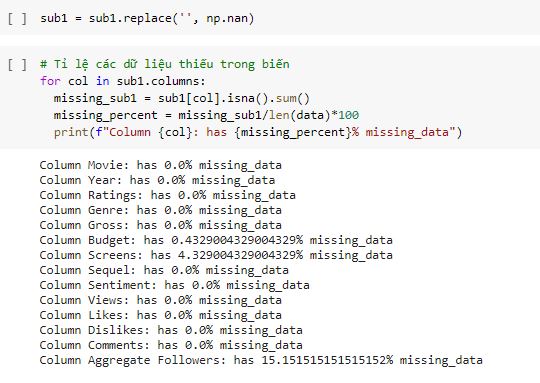
1. Thông tin tập dữ liệu

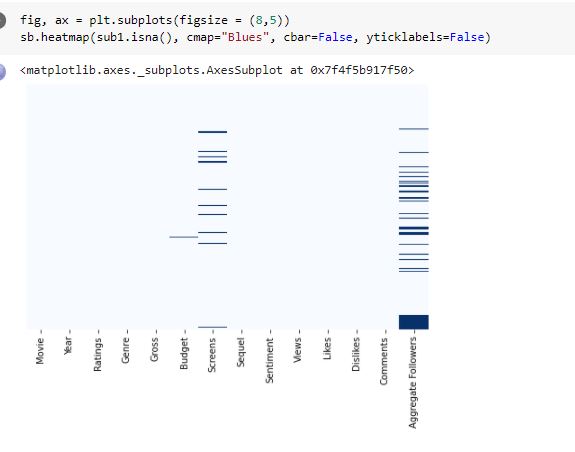
 Xem thông tin các cột của tập dữ liệu bằng phương thức .info(). Tập dữ liệu “Conventional and Social Media Movies” với 231 dòng, 14 cột. Trong đó, ở cột Budget có 1 dòng null, cột Screens có 10 dòng null, cột Aggregate Follower có 35 dòng null.

Hình 1: Thông tin dữ liệu

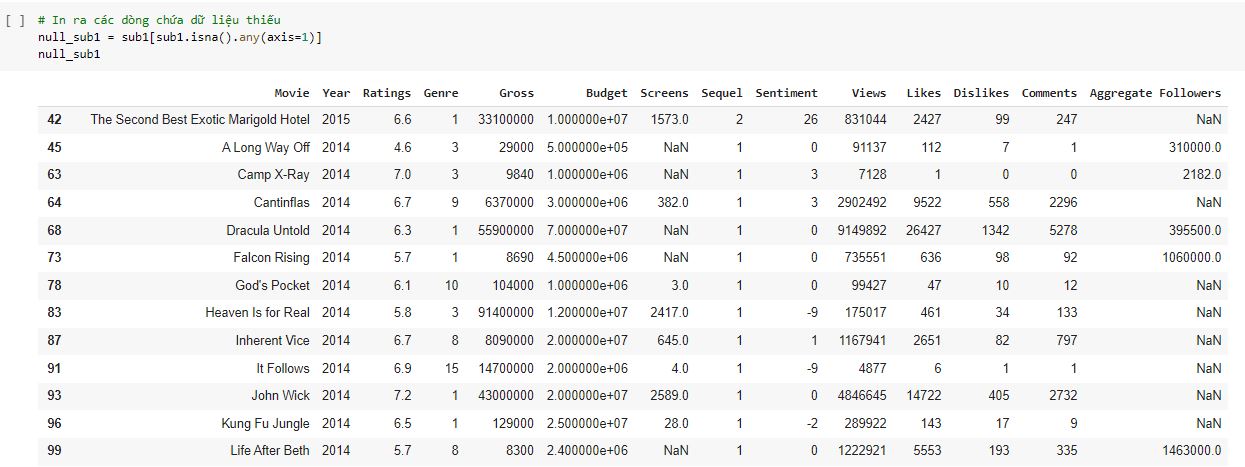
1. Loại bỏ dữ liệu thiếu

Ta có thể tính được tỉ lệ các dữ liệu thiếu trong biến. Kết quả cho thấy trong tập dữ liệu cột Aggregate Followers có tỉ lệ dữ liệu thiếu rất cao có thể ảnh hưởng đến các phân tích sau này.

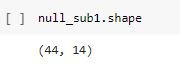
Hình 2: Tỉ lệ dữ liệu thiếu của biến

 Hãy xem thử các bộ phim thiếu cột Aggregate Followers có dữ liệu như thế nào với trực quan hóa dữ liệu thiếu bằng biểu đồ nhiệt.

Hình 3: Trực quan hóa dữ liệu thiếu bằng biểu đồ nhiệt



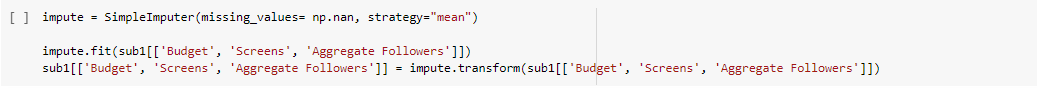
Hình 4: In ra các dòng dữ liệu thiếu

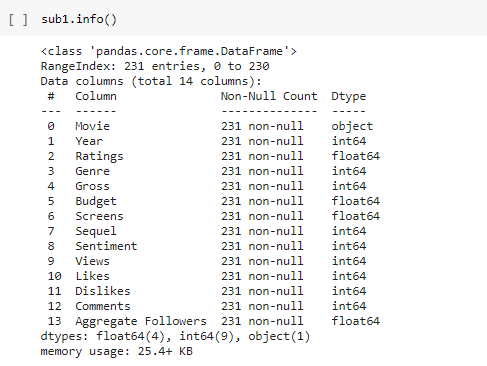
 Xem tổng số lượng dòng chứa dữ liệu thiếu. Kết quả cho thấy tập dữ liệu có 231 dòng mà có đến 44 dòng chứa dữ liệu thiếu.

Hình 5: Số dòng dữ liệu thiếu

 Để xử lý các dữ liệu thiếu thì phương án đề ra là tạo giá trị mean (trung bình mẫu) cho các cột để lấp vào các giá trị thiếu này. Đầu tiên, import SimpleImputer từ thư viện sklearn.impute.

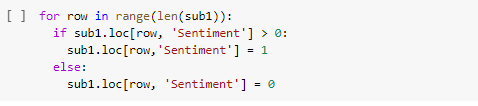
Hình 6: Import lớp SimpleImputer từ thư viện sklearn.impute

*Hình 7: Tạo giá trị mean*

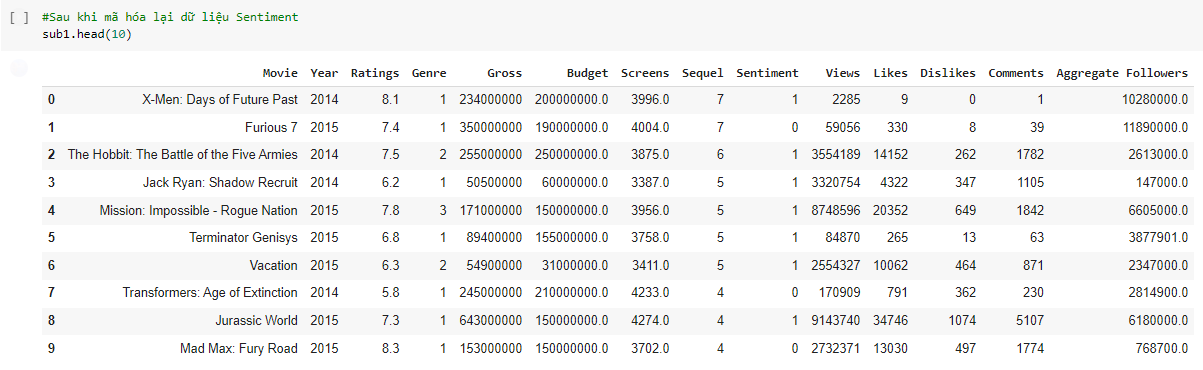
 Xem lại thông tin dữ liệu thì thấy không còn các dòng dữ liệu thiếu nữa.

Hình 8: Thông tin dữ liệu sau khi cập nhật

1. Mã hóa dữ liệu

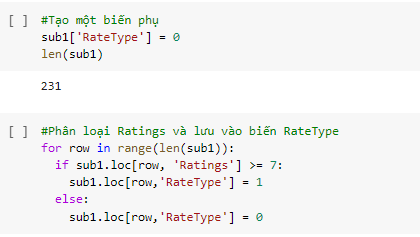
Chuyển đổi dữ liệu biến Sentiment (điểm tình cảm tích cực hay tiêu cực) thành biến định tính: Với giá trị 0 là tiêu cực (Sentiment <= 0 ), giá trị 1 là tích cực (Sentiment > 0).

Hình 9: Mã hóa dữ liệu

Hình 10: Kết quả sau khi mã hóa

1. Nhóm các giá trị và tạo biến phụ

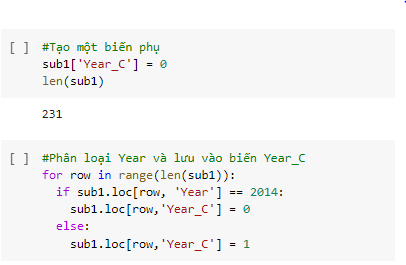
Từ biến Ratings chúng ta phân loại thành 2 nhóm và đồng thời tạo biến phụ RateType để chứa thuộc tính phân loại này :

* Những bộ phim có Ratings >=7
* **Những bộ phim có Ratings < 7

Hình 11: Tạo biến phụ cho Ratings

Tiếp tục với biến Year chúng ta phân loại thành 2 nhóm và đồng thời tạo biến phụ Year\_C để chứa các thuộc tính phân loại này:

* 1. Những bộ phim có Year = 2014
* 2. Những bộ phim có Year = 2015

*Hình 12: Tạo biến phụ cho Year*

# **CHƯƠNG 3: KIỂM ĐỊNH GIẢ THUYẾT**

1. Phân tích phương sai ANOVA

Kiểm định giả thuyết bằng ANOVA được sử dụng trong trường hợp:

+ Biến phản hồi: Định lượng.

+ Biến giải thích: Phân loại.

1. Phân tích ANOVA cho biến giải thích có 2 loại

Câu hỏi nghiên cứu: Liệu năm sản xuất bộ phim có ảnh hưởng đến tổng doanh thu hay không?

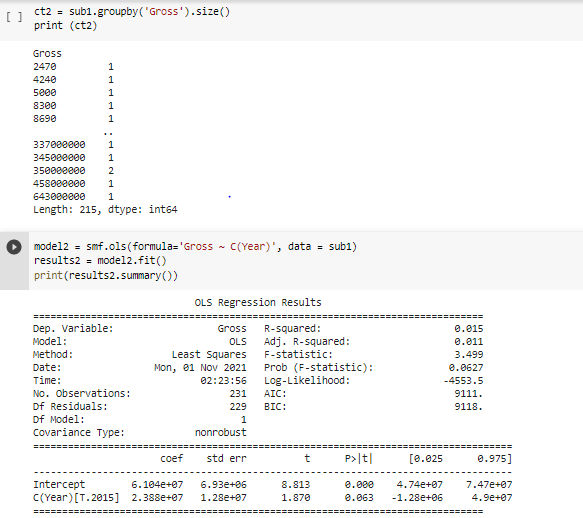
- 2 biến được sử dụng cho câu hỏi nghiên cứu trên là Year (Năm sản xuất) và Gross (Tổng doanh thu).

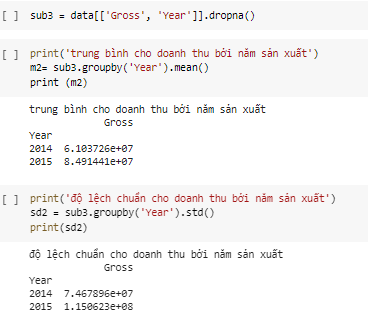
+ Biến phản hồi: Gross

+ Biến giải thích: Year

- Giả thuyết vô hiệu: Năm ra mắt bộ phim không có ảnh hưởng gì tới tổng doanh thu.

- Giả thuyết đối lập: Năm ra mắt bộ phim có ảnh hưởng tới tổng doanh thu.



Hình 13: Phân tích ANOVA cho biến giải thích có 2 loại

*Hình 14: Tính trung bình và độ lệch chuẩn*

Kết quả: Sau khi phân tích ANOVA, ta có F = 3.499, giá trị p = 0.0627 (với mức ý nghĩa là 0.05). Điều đó cho thấy không thể bác bỏ giả thuyết vô hiệu cho rằng: Năm sản xuất phim không ảnh hưởng đến tổng doanh thu. Giá trị p khoảng 0.06 cho thấy nếu chúng ta cho rằng có mối liên quan giữa năm sản xuất và tổng doanh thu thì sẽ sai 6 trên 100 trường hợp quan sát. Số lượng sai khá lớn nên chúng ta bác bỏ giả thuyết vô hiệu.

1. Phân tích ANOVA cho biến giải thích nhiều hơn 2 loại.

Câu hỏi nghiên cứu: Thể loại phim có ảnh hưởng đến tổng doanh thu hay không?

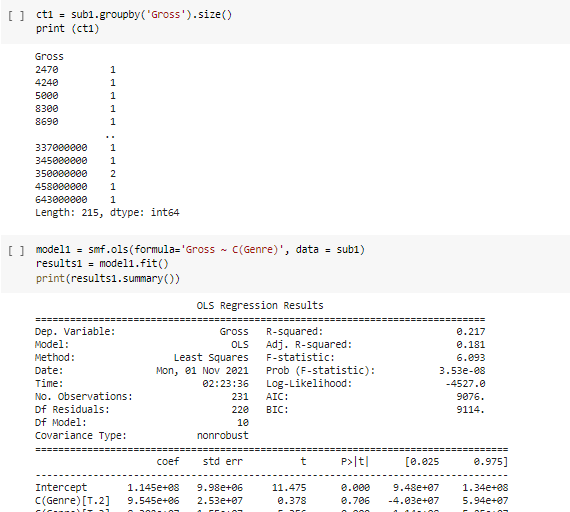
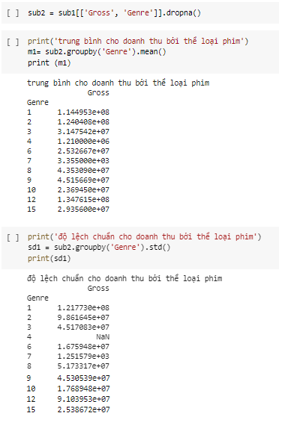
- 2 biến được sử dụng cho câu hỏi nghiên cứu trên là: Genre (Thể loại), Gross (Tổng doanh thu).

+ Biến phản hồi: Gross

+ Biến giải thích: Genre

- Giả thuyết vô hiệu: Thể loại phim không ảnh hưởng tới tổng doanh thu.

- Giả thuyết đối lập: : Thể loại phim có ảnh hưởng tới tổng doanh thu.

Hình 15: Phân tích ANOVA cho biến giải thích nhiều hơn 2 loại

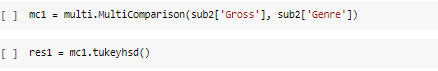
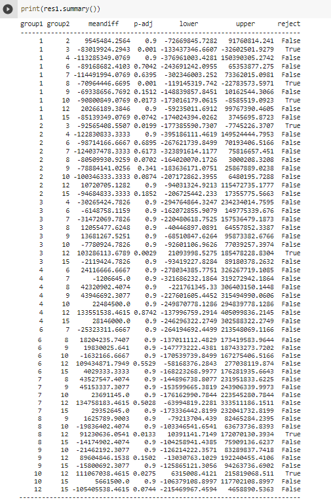
Hình 16: Tính trung bình và độ lệch chuẩn

Kết quả: Sau khi phân tích ANOVA cho vấn đề trên, ta thu được F = 6,093, giá trị p = 3.53e – 08 (với mức ý nghĩa là 0.05). Vì thế có thể bác bỏ giả thuyết vô hiệu cho rằng: Thể loại phim không có ảnh hưởng đến tổng doanh thu. Số lượng sai rất thấp nên chúng ta có thể bác bỏ giả thuyết vô hiệu.

1. Phân tích sâu ANOVA

Cuối cùng để thấy rõ hơn mối quan hệ giữa 2 biến thể loại và tổng doanh thu ta dùng phân tích sâu trong ANOVA.

Phân tích sâu bằng hàm tukeyhsd().

Hình 17: Hàm tukeyhsd()

Hình 18: Kết quả phân tích sâu ANOVA

Từ phân tích sâu,

1. Phân tích Chi – square

Kiểm định giả thuyết bằng Chi – square sử dụng trong trường hợp:

+ Biến phản hồi: Phân loại.

+ Biến giải thích: Phân loại hoặc định lượng.

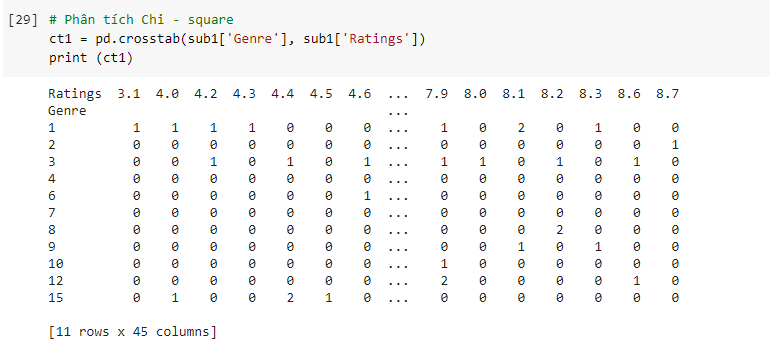
Câu hỏi nghiên cứu: Liệu có mối quan hệ nào giữa xếp hạng và thế loại của bộ phim hay không?

Biến được sử dụng cho câu hỏi trên là:

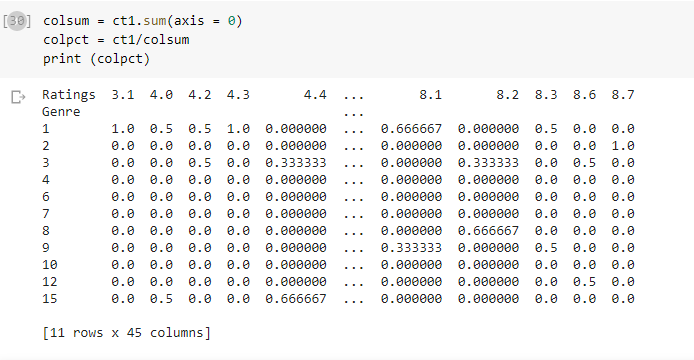
+ Biến phản hồi: Genre

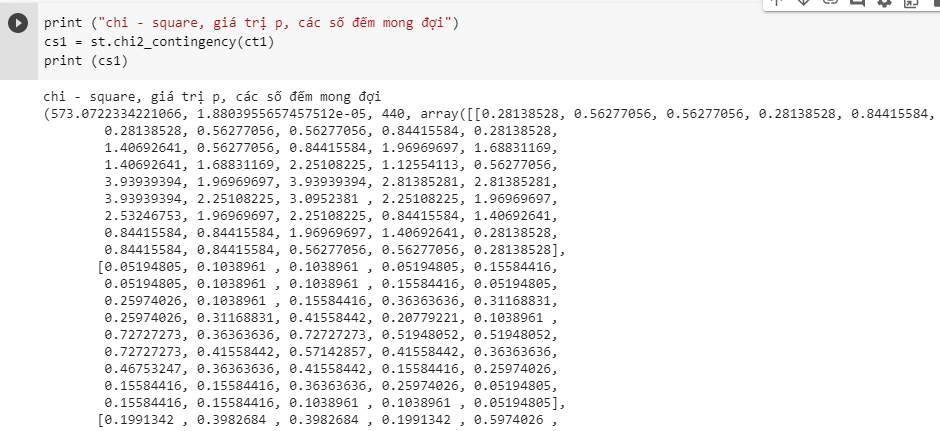
+ Biến giải thích: Ratings

- Giả thuyết vô hiệu: Giữa xếp hạng và thể loại phim không có mối quan hệ nào hết.

 - Giả thuyết đối lập: Giữa xếp hạng và thể loại phim có mối quan hệ với nhau.

Hình 19: Dùng hàm pandas.crosstab

Hình 20: Tính phần trăm cột

Hình 21: Tính Chi - square

Kết luận: Ta thấy giá trị p = 1.88e – 05 (với mức ý nghĩa là 0.05)

1. Tương quan Pearson

Tương quan Pearson được sử dụng trong trường hợp:

+ Biến phản hồi: Định lượng.

+ Biến giải thích: Định lượng.

Câu hỏi nghiên cứu: Có mối liên hệ nào giữa số lượng bình chọn và tổng doanh thu hay không?

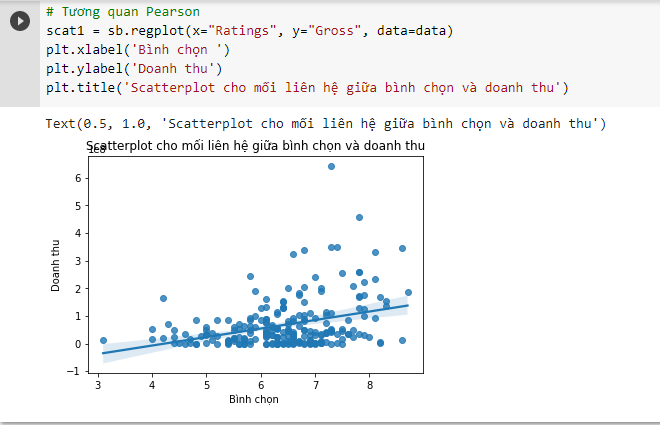
+ Biến phản hồi: Ratings

+ Biến giải thích: Gross

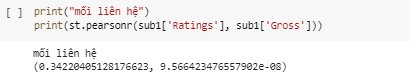
- Giả thuyết vô hiệu: Giữa 2 đại lượng không có mối liên hệ.

- Giả thuyết đối lập: Giữa số lượng bình chọn và tổng doanh thu có mối quan hệ với nhau.

Đồ thị để phân tích giữa 2 biến định lượng là đồ thị phân tán để thấy được mối quan hệ giữa 2 biến với nhau.



Hình 22: Đồ thị phân tán

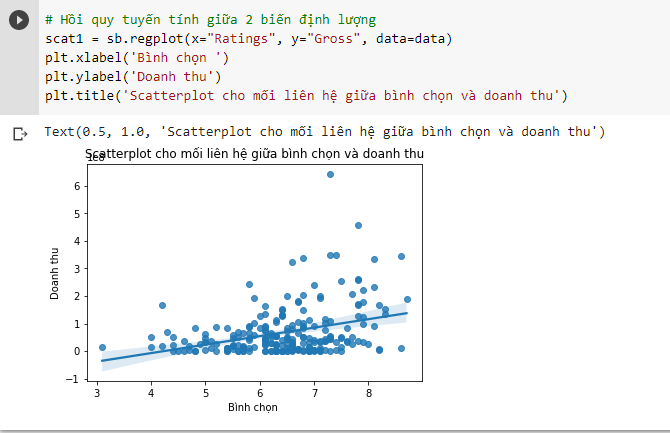
**** Nhìn vào đồ thị có thể nhận xét rằng mối quan hệ giữa bình chọn và doanh thu là dạng tuyến tính dương. Nghĩa là nếu lượng bình chọn tăng thì doanh thu cũng tăng.

Hình 23: Tính hệ số tương quan

Hệ số tương quan Pearson = 0.34 thấp so với 1, chứng tỏ sự tương quan giữa 2 biến khá yếu. Hệ số p = 9.56e – 08 (với mức ý nghĩa là 0.05) cho thấy mối quan hệ này có ý nghĩa.

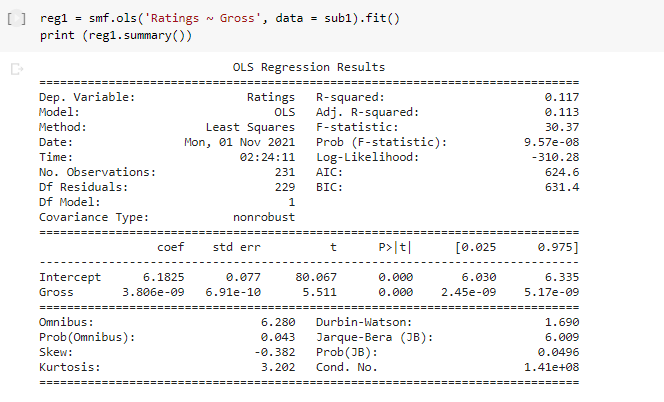
# **CHƯƠNG 4: PHÂN TÍCH HỒI QUY**

1. Hồi quy tuyến tính giữa 2 biến định lượng

 Cùng xem xét mối quan hệ tuyến tính giữa 2 biến định lượng Ratings (Lượng bình chọn) và Gross (Tổng doanh thu). Nếu bộ phim có lượng bình chọn cao thì lợi nhuận thu vào có đáng kể không?

Hình 24: Biểu đồ phân tán 2 biến định lượng

Từ biểu đồ phân tán, ta đánh giá được lượng bình chọn tỉ lệ thuận với doanh thu ở mức khá thấp. Có một số bộ phim có doanh thu cao nhưng lượt bình chọn chỉ ở mức trung bình so với các bộ phim khác. Tuy nhiên, cũng dễ thấy rằng những bộ phim có lượt bình chọn cao thì doanh thu cũng ở mức khá cao.

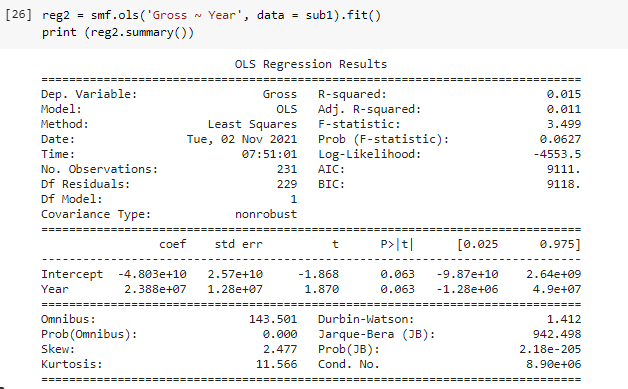
Hình 25: Mô hình hồi quy tuyến tính 2 biến định lượng

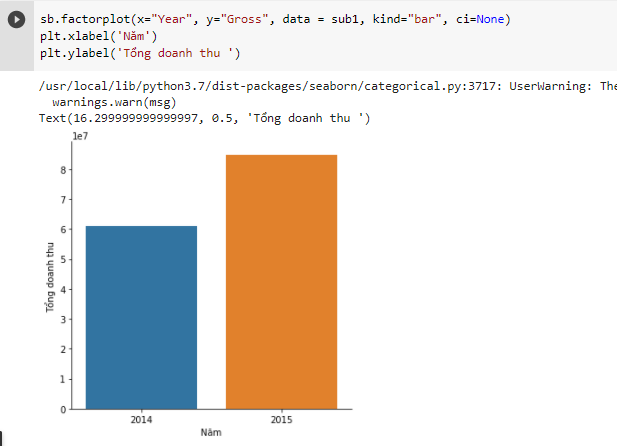
Ta có, p = 9.57e – 08 (với mức ý nghĩa là 0.05) nên chúng ta có thể khẳng định lượng bình chọn không làm ảnh hưởng đến doanh thu của bộ phim. Thu được phương trình hồi quy:

Ratings = 6.1825 + 3.806e – 09 \* Gross

1. Hồi quy tuyến tính giữa biến định lượng và phân loại

Xem xét mối quan hệ tuyến tính giữa Year (Năm sản xuất) và Gross (Tổng doanh thu), trong đó Year là 1 biến số giải thích phân loại (2014 và 2015) và Gross là biến phản hồi định lượng. Câu hỏi nghiên cứu là: Năm sản xuất có liên quan đến doanh thu của bộ phim tăng lên hay không?

Hình 26: Mô hình hồi quy tuyến tính giữa biến phân loại và định lượng

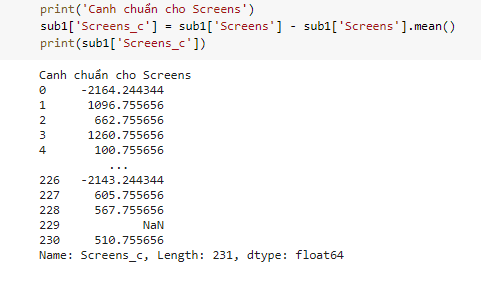
 Từ kết quả trên, ta thấy rằng giá trị p = 0.0627 (với mức ý nghĩa là 0.05) nên khẳng định năm sản xuất không ảnh hưởng đến doanh thu bộ phim tăng lên. Do đó, không thu được phương trình hồi quy. Biểu đồ tương ứng bên dưới sẽ cho ta thấy một cách trực quan hơn.

Hình 27: Biểu đồ cho cái nhìn trực quan cho mô hình

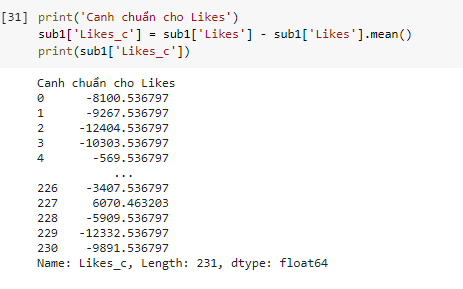
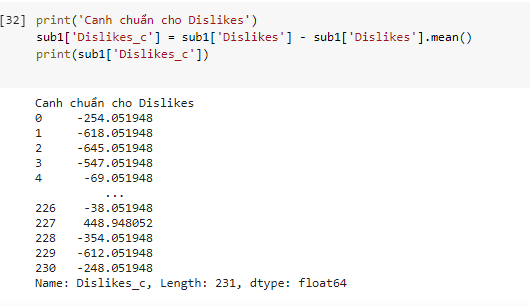
1. Hồi quy đa biến

Câu hỏi đặt ra ở mô hình tuyến tính trên: Nếu bộ phim có lượng bình chọn cao thì doanh thu có đáng kể không?. Với nghiên cứu trên thì ta đã biết đáp án là có. Cũng như các ngành giải trí khác thì doanh thu của phim điện ảnh cũng chịu ảnh hưởng từ nhiều tác nhân chủ quan và khách quan khác nhau. Tất nhiên, tập dữ liệu này cũng sẽ chịu ảnh hưởng từ nhiều biến ẩn khác, gây ảnh hưởng không mong muốn đến kết quả của nghiên cứu.

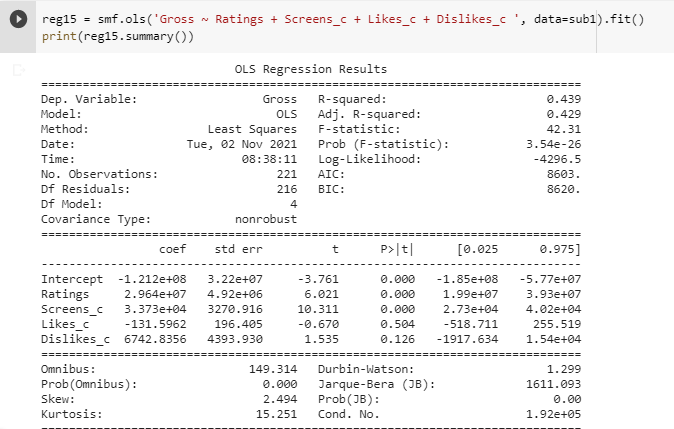
Ta đã có bằng chứng cho thấy rằng lượng bình chọn có liên quan đáng kể đến tổng doanh thu. Tất nhiên, một vài yếu tố khác có khả năng ảnh hưởng đến tổng doanh thu là số lượng rạp chiếu, số lượt yêu thích và số lượng không thích.

 Để đánh giá điều này đúng không, ta thêm các biến số lượng rạp chiếu (Screens), số lượt yêu thích (Likes), số lượt không thích (Dislikes).

Hình 28: Canh chuẩn cho Screens

Hình 29: Canh chuẩn cho Likes

Hình 30: Canh chuẩn cho Dislikes



Hình 31: Mô hình hồi quy đa biến

Kết quả từ mô hình hồi quy, ta thấy rằng giá trị p của các biến Ratings, Screens có giá trị là 0 nhỏ hơn mức ý nghĩa là 0.05 nên các biến này có tác động đến biến phản hồi Gross. Còn các biến Likes, Dislikes có giá trị p lần lượt là 0.504 và 0.126 lớn hơn 0.05 nên các biến này không có tác động đến biến phản hồi Gross. Vì thế có thể kết luận trong các yếu tố có khả năng ảnh hưởng đến doanh thu của bộ phim mà chúng ta quan tâm thì lượng bình chọn và số lượng rạp chiếu

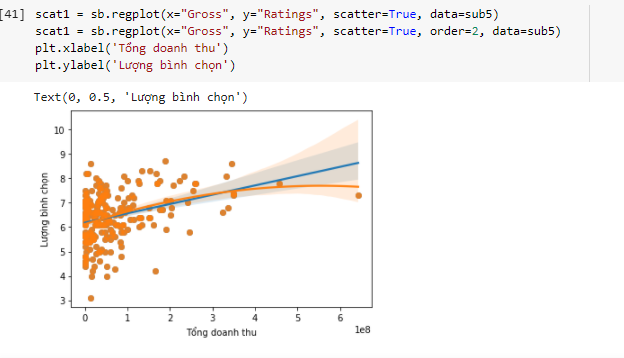
quyết định giá trị doanh thu của bộ phim. Đồng thời, ta thu được phương trình hồi quy sau:

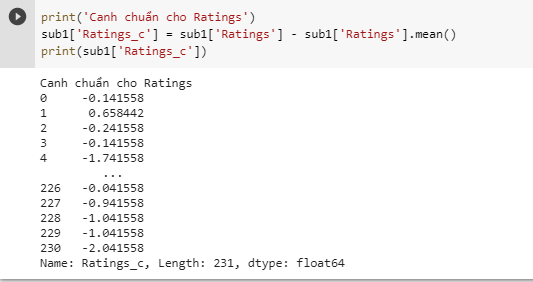
Gross = 2.964e + 07 \* Ratings + 3.373e + 04 \* Screens\_c - 1.212e+08

1. Hồi quy đa thức

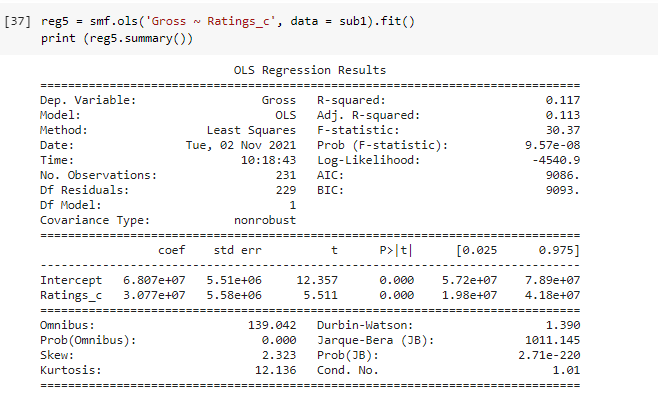
Tiếp tục với câu hỏi nghiên cứu quen thuộc: Xem xét giữa lượt bình chọn và tổng doanh thu có ảnh hưởng với nhau hay không?

Vẽ đồ thị phân tán biểu diễn cho 2 biến với 1 đường đồ thị tuyến tính và 1 đường đồ thị bậc 2.

Hình 32: Đồ thị phân tán biểu diễn 2 biến

 Mặc dù, chúng ta đã vẽ được đồ thị nhưng vẫn chưa thấy rõ lắm về mối liên hệ bậc 2 có tốt hay không. Tiếp theo, chúng ta dẽ thực hiện canh chuẩn dữ liệu cho biến Ratings.

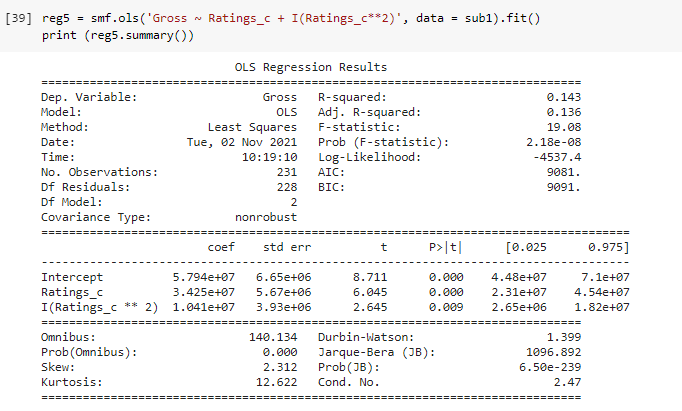
Hình 33: Canh chuẩn dữ liệu cho Ratings



Hình 34: Mô hình hồi quy bậc 1

Kết quả cho ta thấy giá trị p của biến Ratings\_c p = 0 nhỏ hơn 0.05 tức là có ý nghĩa, nên giữa doanh thu và lượt bình chọn không có ảnh hưởng với nhau. Tuy nhiên, hãy quan tâm đến 1 giá trị khác là R – squared = 0.117, điều này cho thấy tương quan Pearsoon của mối quan hệ giữa 2 biến Gross và Ratings rất thấp. Vì thế chưa thể khẳng định chắc chắn mối quan hệ giữa 2 biến này.

Từ vấn đề trên, ta tiếp tục với thử với một mô hình bậc cao hơn cho câu hỏi nghiên cứu này.

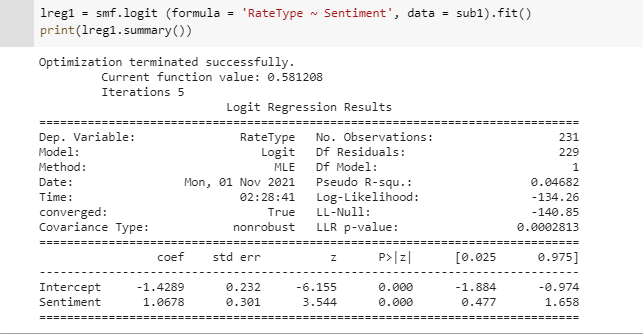
Hình 35: Mô hình hồi quy bậc 2

Ta thấy giá trị p của cả 2 biến bậc 1 và bậc 2 đều có ý nghĩa vì nhỏ hơn 0.05. Đặc biệt, giá trị p của biến bậc 2 chứng minh cho mô hình đường cong quan sát được trong mô hình phân tán có ý nghĩa. Ngoài ra, giá trị R – square = 0.143 lớn hơn R – square bậc 1 (0.117) chứng tỏ mối liên hệ giữa 2 biến trong mô hình bậc 2 tốt hơn so với mô hình bậc 1. Kết luận: Mô hình đường cong bậc 2 phù hợp hơn so với mô hình đường thẳng.

1. Hồi quy Logistic
2. Hồi quy Logistic đơn biến

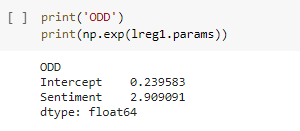
Xem xét ảnh hưởng của lượt bình chọn với điểm tình cảm. Sử dụng 2 biến trong quá trình phân tích: RateType (Biến phụ của Ratings) và Sentiment (Điểm tình cảm) đã được mã hóa thành 0, 1.

+ Biến giải thích: Sentiment

 + Biến phản hồi: RateType

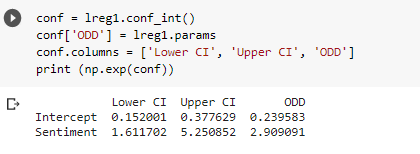
Hình 36: Mô hình hồi quy Logistic đơn biến

Kết quả: Số lượng quan sát là 231. Giá trị p của biến Sentiment là 0 rất nhỏ, nhưng trong mô hình hồi quy Logistic thì phương trình hồi quy tuyến tính không còn phù hợp nữa. Và giá trị của biến phản hồi chỉ nhận giá trị là 0 và 1. Do đó, ta tiếp tục phân tích mô hình này.

 Tính ODD là tỷ số giữa xác suất xuất hiện biến cố và xác suất không xuất hiện biến cố.

Hình 37: Tính ODD

Kết luận: Những bộ phim có điểm tình cảm tích cực thì lượng bình chọn cao hơn 2.9 lần so với những bộ phim có điểm tình cảm tiêu cực.

 Để có thể kết luận cho tổng thể thì ta tiếp tục dựa vào đại lượng khoảng tin cây.

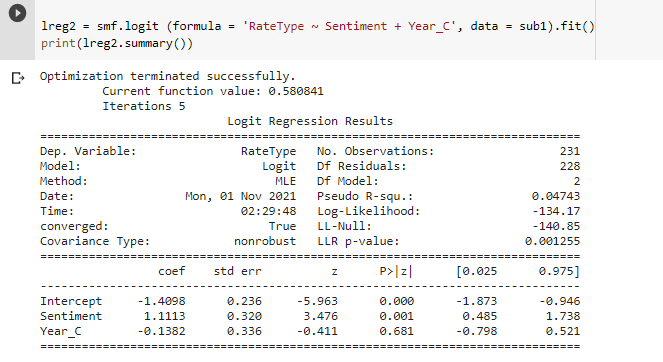
Hình 38: Lấy khoảng tin cậy

Kết luận: Có 95% khả năng ODD của tổng thể nằm trong khoảng 1.6 đến 5.25. Do đó, những bộ phim có điểm tình cảm tích cực thì lượng bình chọn cao hơn từ 1.6 đến 5.25 lần so với những bộ phim có điểm tình cảm tiêu cực.

1. Hồi quy Logistic đa biến

Xem xét ảnh hưởng của lượt bình chọn với điểm tình cảm và năm sản xuất. Sử dụng 3 biến trong quá trình phân tích: RateType (Biến phụ của Ratings), Sentiment (Điểm tình cảm) đã được mã hóa thành 0, 1 và Year\_C (Biến phụ của Year)

+ Biến giải thích: Sentiment, Year\_C

 + Biến phản hồi: RateType

Hình 39: Mô hình hồi quy Logistic đa biến

# **Chương 5: Cây quyết định**

Việc đánh giá một bộ phim phụ thuộc vào nhiều yếu tố. Hơn hết, thái độ tích cực của người xem có thể giúp bộ phim nổi tiếng hơn, tăng doanh thu và lượng người theo dõi hơn, ngược lại, thái độ tiêu cực của người xem gây ra không ít khó khăn cho việc phát triển của bộ phim.

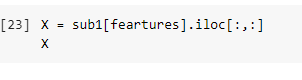
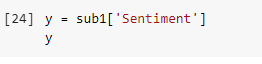
Vậy để đánh giá được tình cảm của người xem là tiêu cực hay tích cực, cần xây dựng một mô hình phân loại tình cảm trên tập dữ liệu CSM.

Trong chương này, ta sử dụng cây quyết định để xây dựng mô hình dự đoán. Tập dữ liệu CSM có khá nhiều các thuộc tính có mối liên hệ xoay quanh thuộc tính tìm cảm(Sentiment), tuy nhiên, sẽ lượt bỏ một số thuộc tính không ảnh hưởng đến thuộc tính Sentiment như Movie, Year, RateType, Year\_C. Cuối cùng ta có được danh sách các thuộc tính feartures như sau:



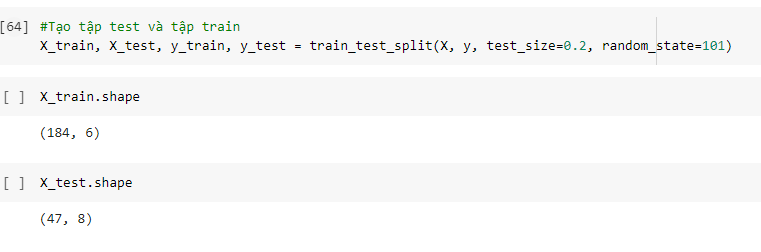
*Hình 40: Các thuộc tính features*

Tách biến không phụ thuộc (independent variable) và biến phụ thuộc (dependent variable) từ tập dữ liệu (dataset).

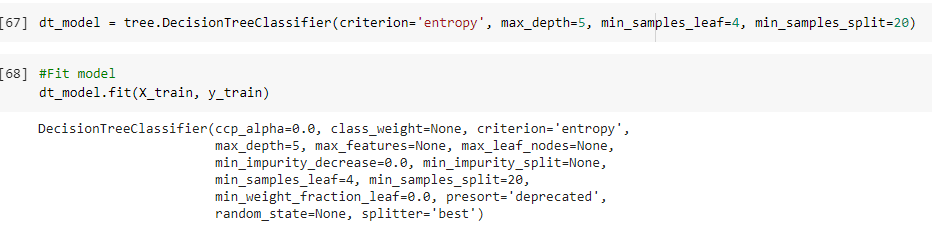
*Hình 41: Tách biến cho mô hình cây quyết định*

Tiếp theo chúng ta sẽ phân chia tập dữ liệu thành hai tập nhỏ hơn: tập dữ liệu huấn luyện (training set) và tập dữ liệu xác nhận (test set) với tỷ lệ 80% và 20%.



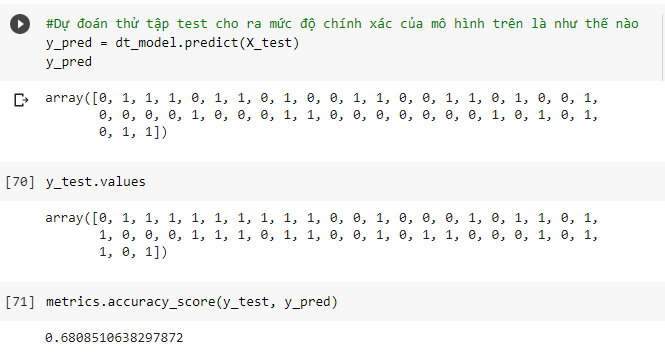
*Hình 42: Chia tập dữ liệu train và test 1*

Sau rất nhiều lần chỉnh sửa và chạy thử với các giá trị khác nhau của các thông số criterion(chỉ số tính độ bất định), max\_depth( độ sâu lớn nhất mà cây có thể phát triển), min\_samples\_split( số cá thể tối thiểu để phân nhánh một nút), min\_samples\_leaf( số cá thể tối thiểu ở lá), tôi thu được các thông số tối ưu nhất có thể cho mô hình của chúng ta như sau:



*Hình 43: Xây dựng mô hình cây quyết định*

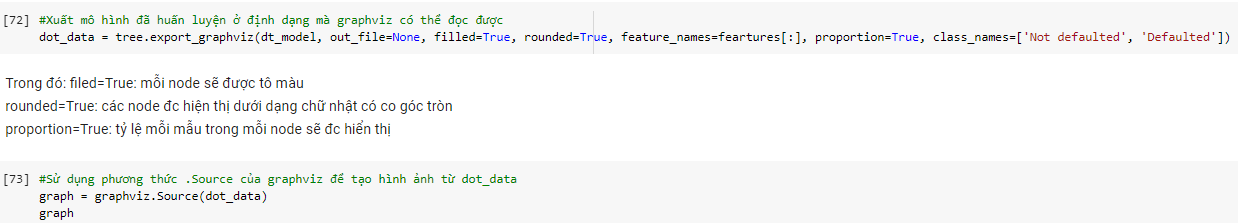
Ở phần trên ta có phân chia tập dữ liệu ra 2 tập train set và test set nhằm thử lại độ chính xác của mô hình. Trong thực tế nếu chỉ dùng một tập dữ liệu để xây dựng cây và cũng dùng tập đó để kiểm thử cây có thể dẫn đến không công tâm lắm hay kết quả thu được quá khớp với mô hình. Vì thế, ta sẽ thử lại mô hình với tập test set:



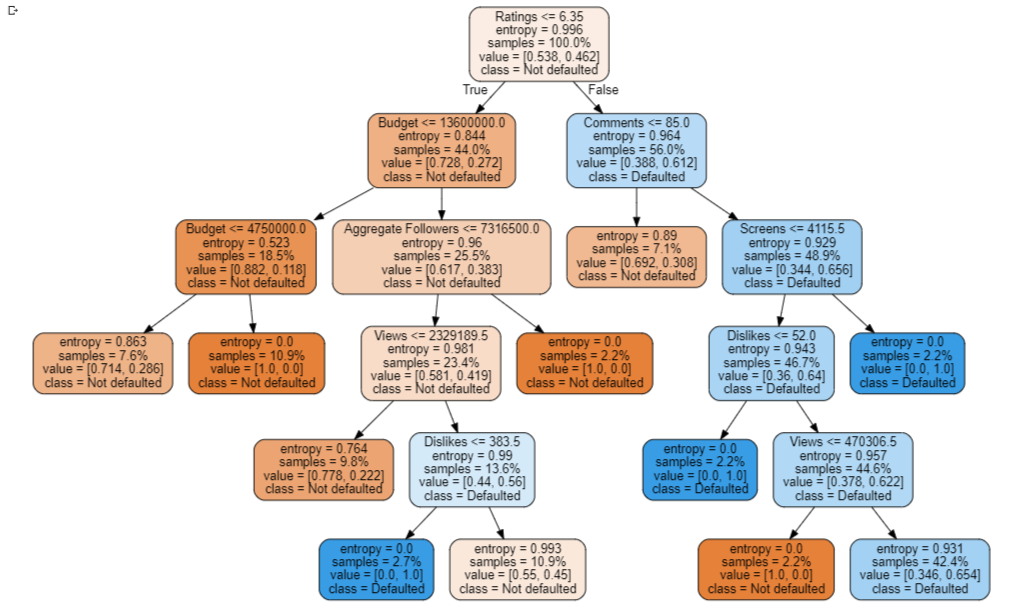
*Hình 44: Kiểm thử mô hình*

Từ kết quả thu được ta xây dựng được mô hình cây quyết định để phân loại tình cảm tích cực hay tiêu cực của người xem với bộ phim với độ chính xác mô hình là 68% khi kiểm thử với tập test. Độ chính xác không cao có thể bị ảnh hưởng bởi số lượng các biến định lượng khá nhiều.

Sử dụng thư viện Graphviz để trực quan hóa mô hình.



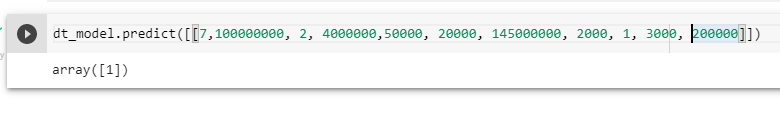
*Hình 45: Trực quan hóa mô hình*



*Hình 46: Cây quyết định*

Hãy cùng làm một mẫu thử thú vị, xem mẫu được đặt ra sẽ được mô hình phân loại vào nhóm có tiềm năng lấy được sự yêu thích của người xem hay không?

Tôi có mẫu như sau: một bộ phim có tỷ lệ đánh giá 7/10, tổng chi phí sản xuất 100000000 đô la, thể loại khoa học viễn tưởng, có 4000000 lượt xem, với 50000 likes, 20000 dislikes, thu về 145000000 đô la, được chiếu tại 2000 rạp, phần đầu tiên, có 3000 bình luận và có 200000 người theo dõi.



Mô hình của chúng ta đã dự đoán mẫu thử trên sẽ có được tình cảm tích cực từ người xem. Theo mô hình cây quyết định như hình 46 và các thông số của mẫu thử ta đã đặt ra. Khi phân loại chỉ có các biến Ratings, Comments, Screens, Dislikes, Views được xét đến, nhánh còn lại của cây bị bỏ qua. Có thể thấy cây quyết định sẽ phân loại dựa trên các luật được xây dựng từ tập dữ liệu mẫu. Vì thế, mẫu thử này cũng có thể sai trong một mô hình với tập dữ liệu train khác.