Universal
Project SmartMovie

16.1.2019

UNIVERSAL + FAB4

Maaike Dokter, Charles Hu, Pournami Krishnan, Carmen Burghardt

Confidential. Do not distribute.
FAB4 LLC 2019

We use movie ratings to provide data-driven insights that help you create more relevant movies, reach more users, and earn more profits.

Agenda

What we can do for you

1. Make suggestions on genres for new movies for targeted age groups
2. Give insights into which genre combination could provide unique opportunities for new movies
3. Provide marketing opportunities by reaching the most active users in their workplace
4. Help people discover new genres
5. Recommend new theme park rides by determining most popular movie for a selected age group

Dataset Analysis

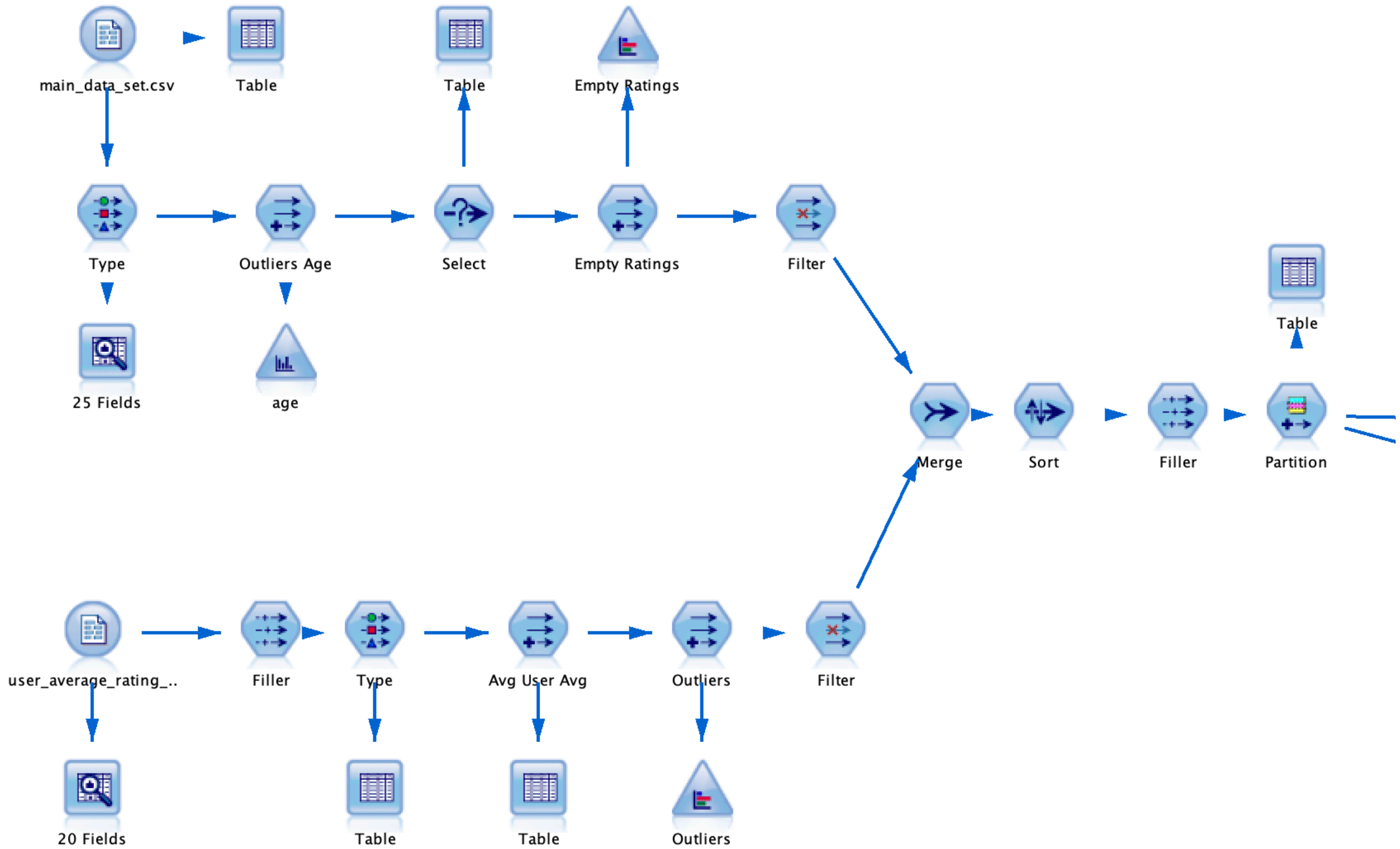
- The data comes from https://movielens.org, which offers non-commercial, personalized movie recommendations.

- Two datasets combined:

    - Main dataset: demographic information on the user, rating given to movie and genre that movie belongs to

    - Average rating per genre: every user has an average rating per genre in this dataset

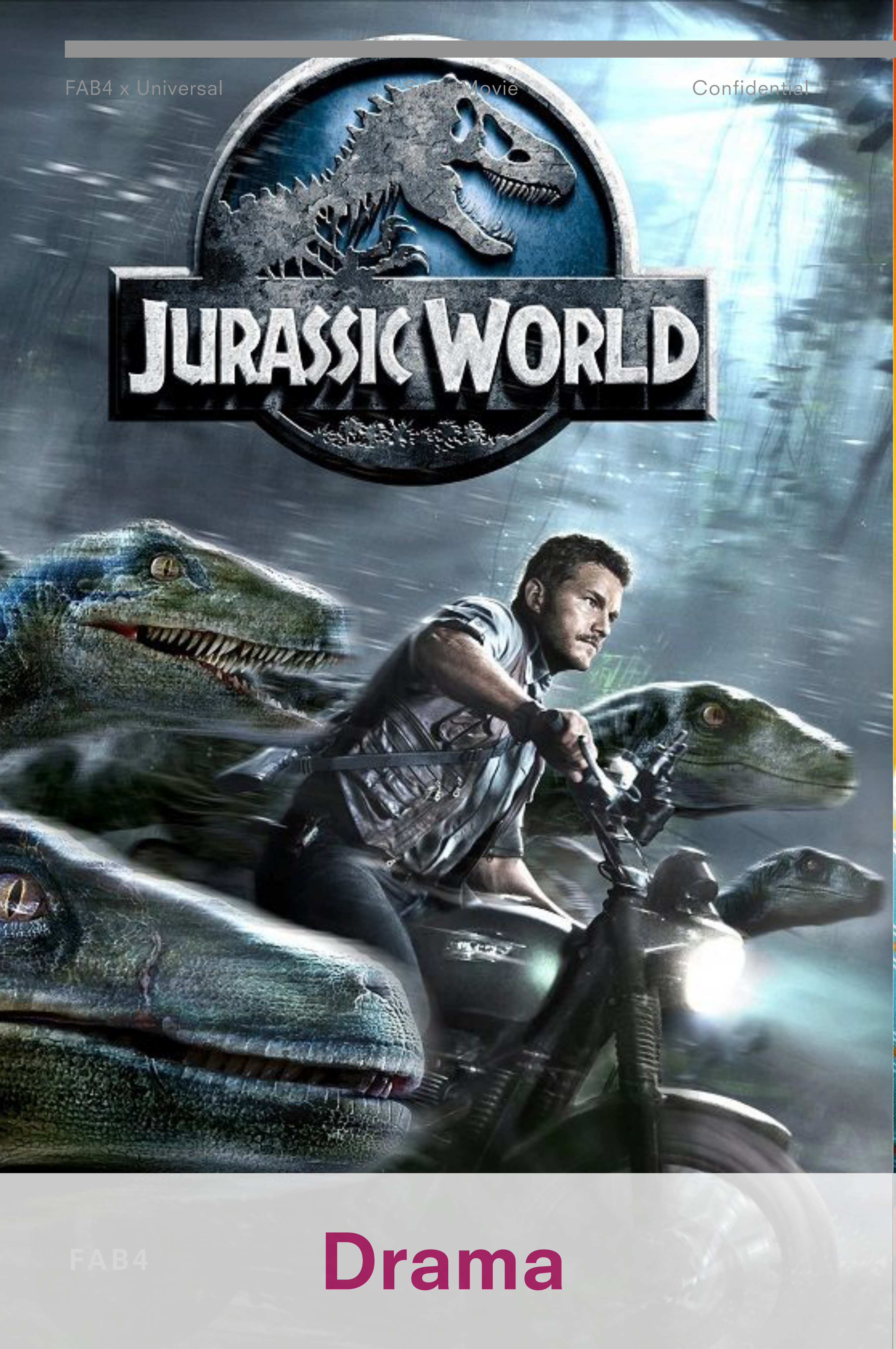
Data Preparation



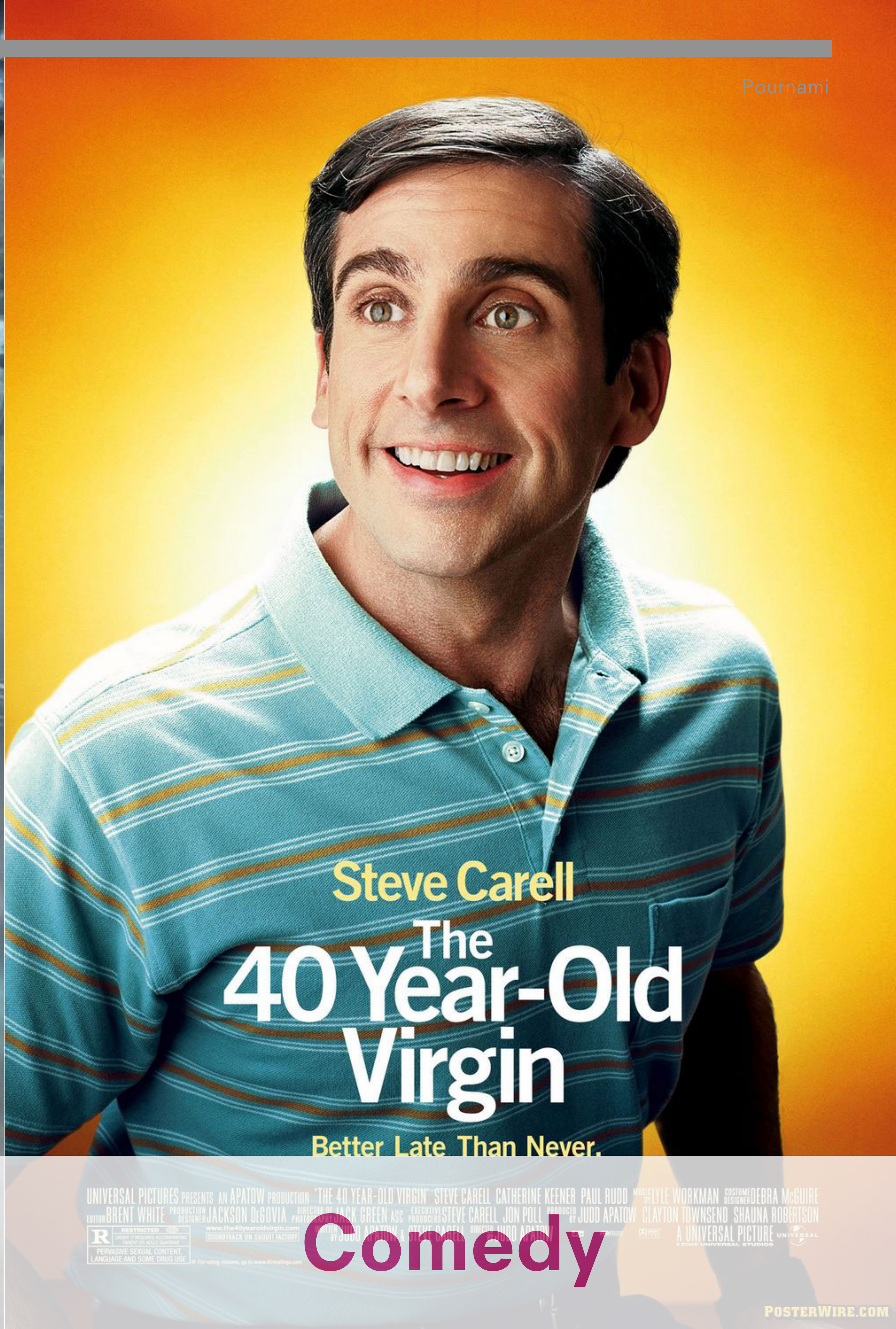
Want to create a movie that can break the box office?

Current popular genre for popular movie watching age group

rating_Mean	rating_Count	Comedy	Drama	Romance
0.834	4044	0	1	0
0.748	2595	1	0	0
0.824	1481	0	1	1
0.713	1457	1	0	1



Drama



Comedy

Drama + Romance

Romance + Comedy

Business goal 01

Make suggestions on genres for new movies for targeted age groups

Understand which is the most popular age group is and see which is their most popular genre of movie: based on this information, movie makers can produce popular genre movies for suitable audiences.

Business goal 02

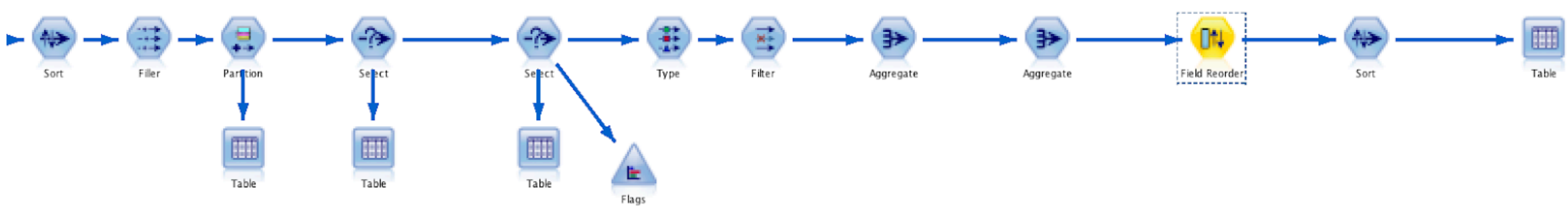
Give insights into which genre combination could provide unique opportunities for new movies

Identify the most popular movie genre combinations and how many movies are already produced, such that a movie producer can create a new movie using optimal genre combinations.

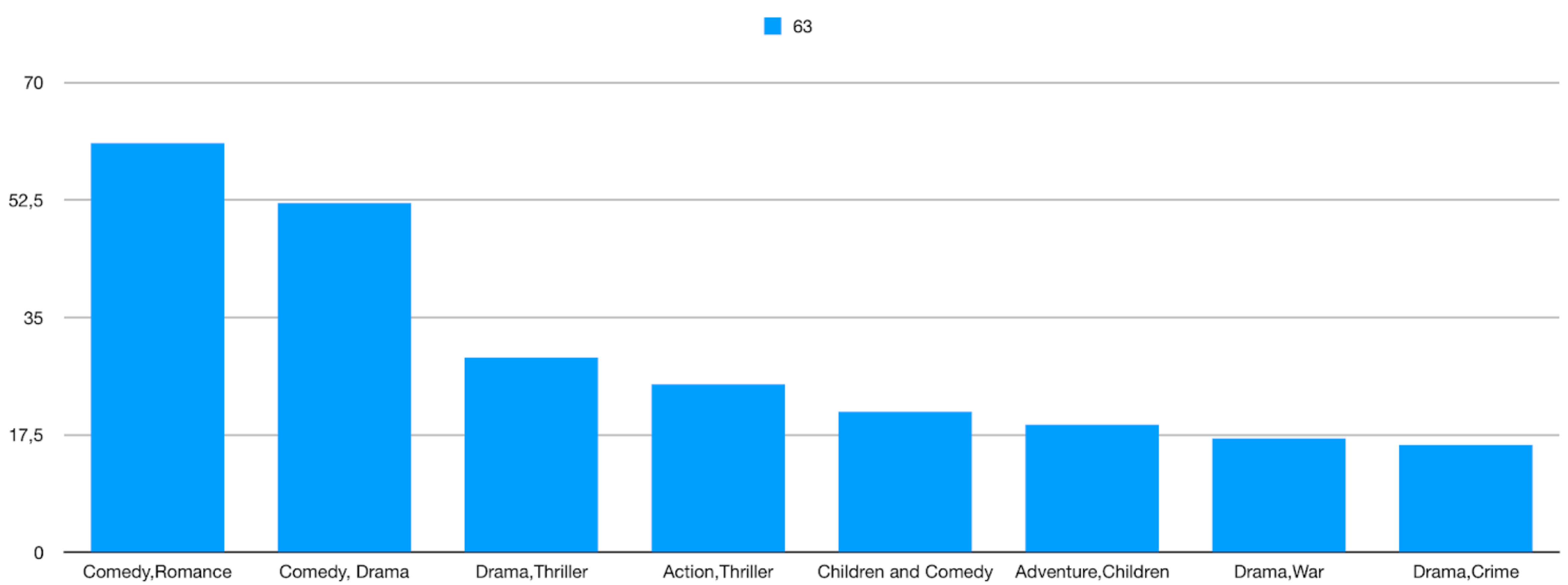
Preparation

Table Annotations

	N Movies	Unkown	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentary	Drama	Fantasy	Film-Noir	Horror	Musical	Mystery	Romance	Sci-Fi	Thriller	War	Western
1	63	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
2	61	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
3	52	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
4	29	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
5	25	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
6	21	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
7	19	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
9	16	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
10	14	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	13	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0
12	12	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
13	11	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
14	11	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0
15	11	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
16	10	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
17	10	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0



Result

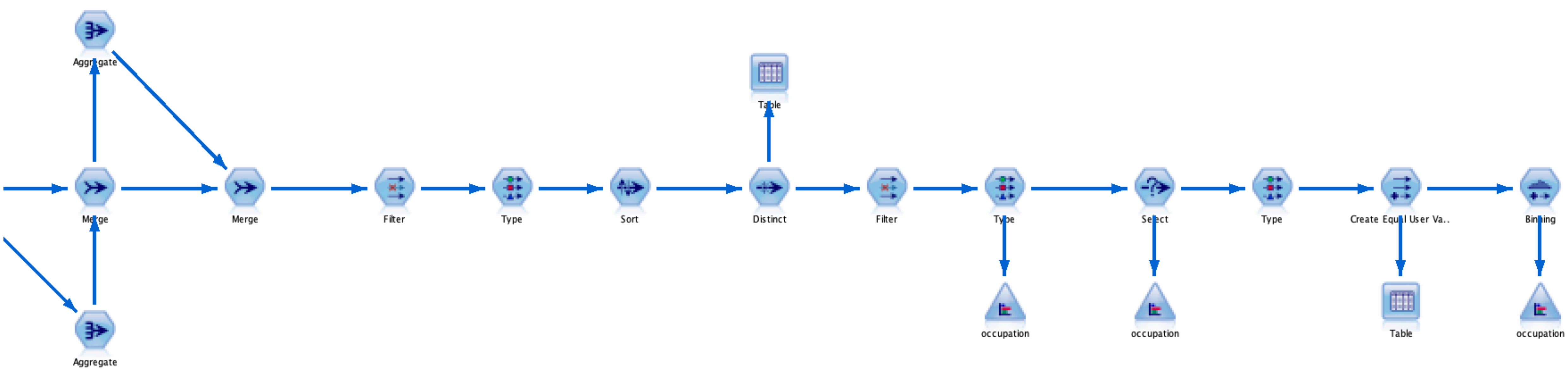

63
70
52,5
35
17,5
0
Comedy,Romance
Comedy, Drama
Drama,Thriller
Action,Thriller
Children and Comedy
Adventure,Children
Drama,War
Drama,Crime

Business goal 03

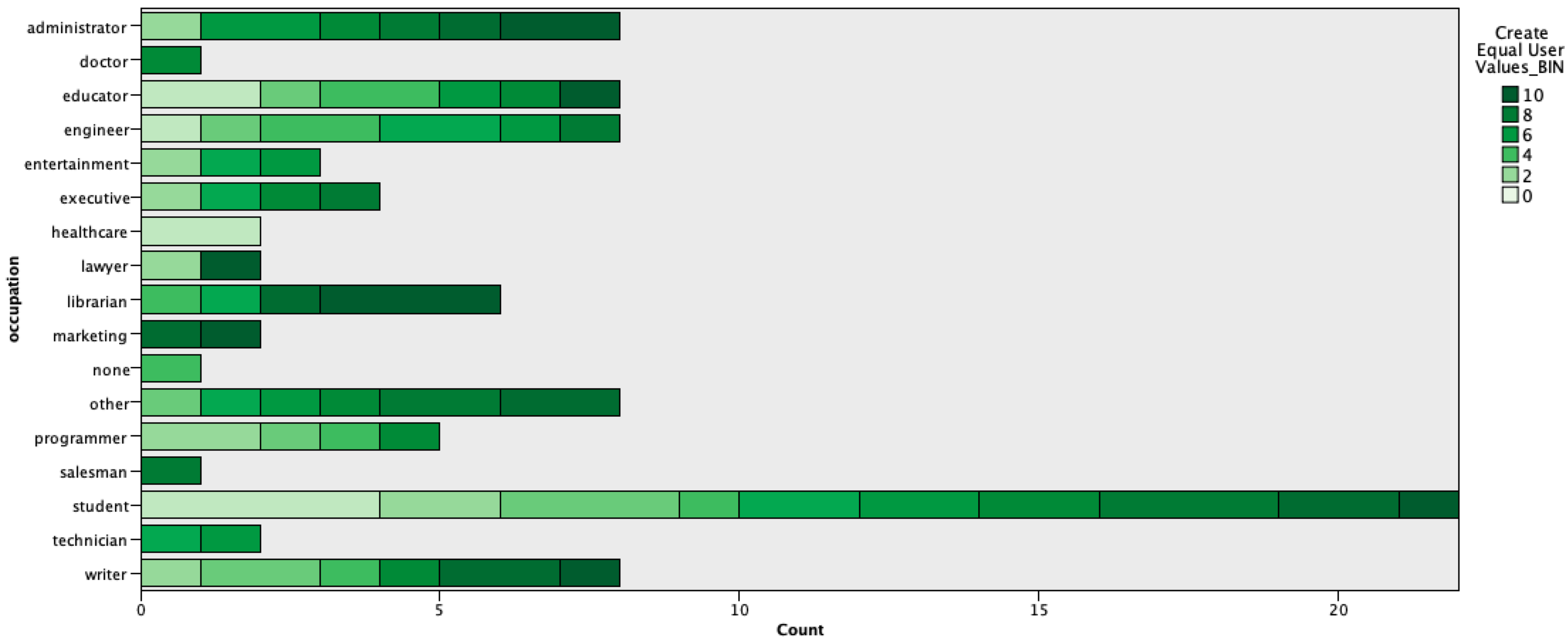
Provide marketing opportunities by reaching the most active users in their workplace

Display the most active users within the most active workplaces

Preparation



Results

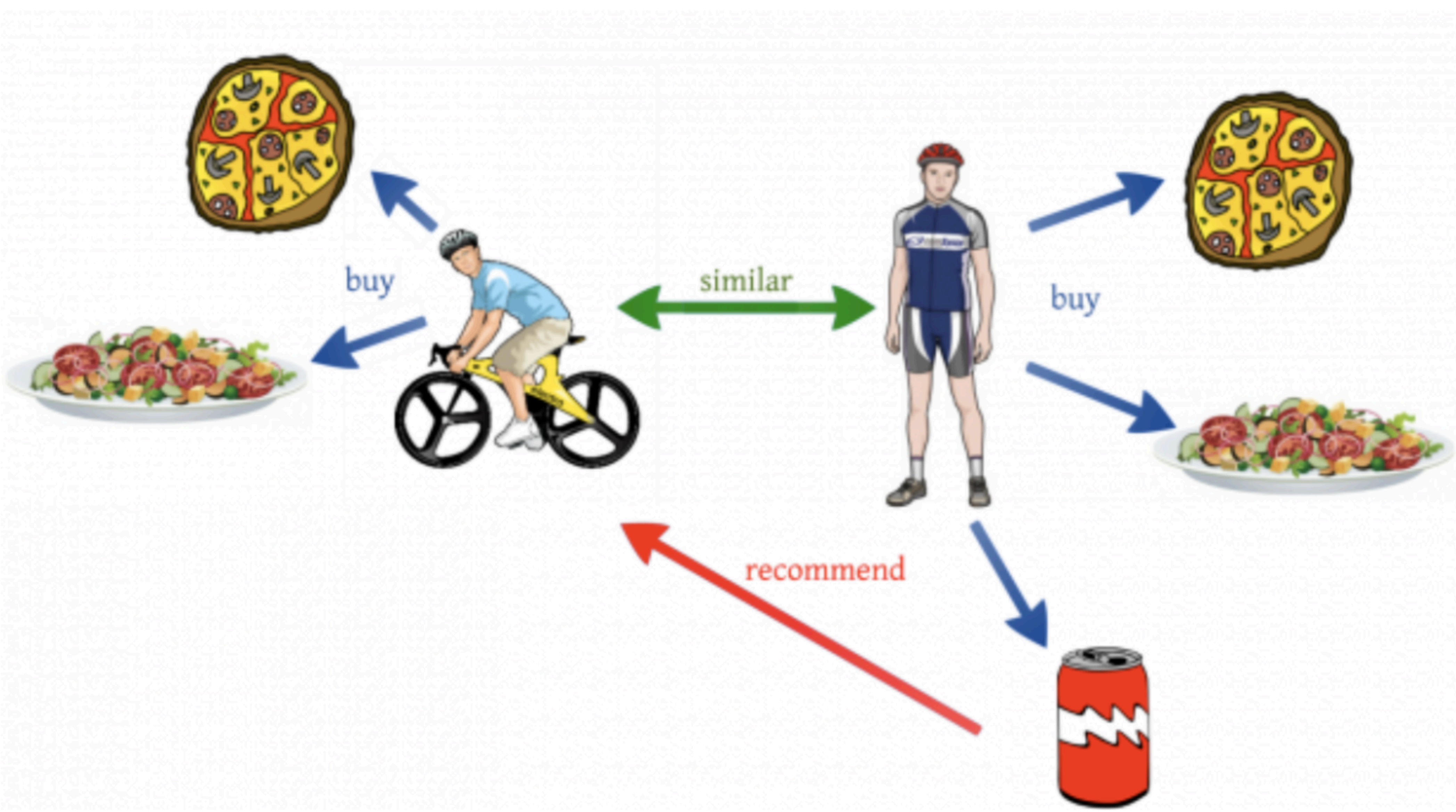

Create Equal User Values_BIN
10
8
6
4
2
0
occupation
administrator
doctor
educator
engineer
entertainment
executive
healthcare
lawyer
librarian
marketing
none
other
programmer
salesman
student
technician
writer
0
5
10
15
20
Count

Business goal 04

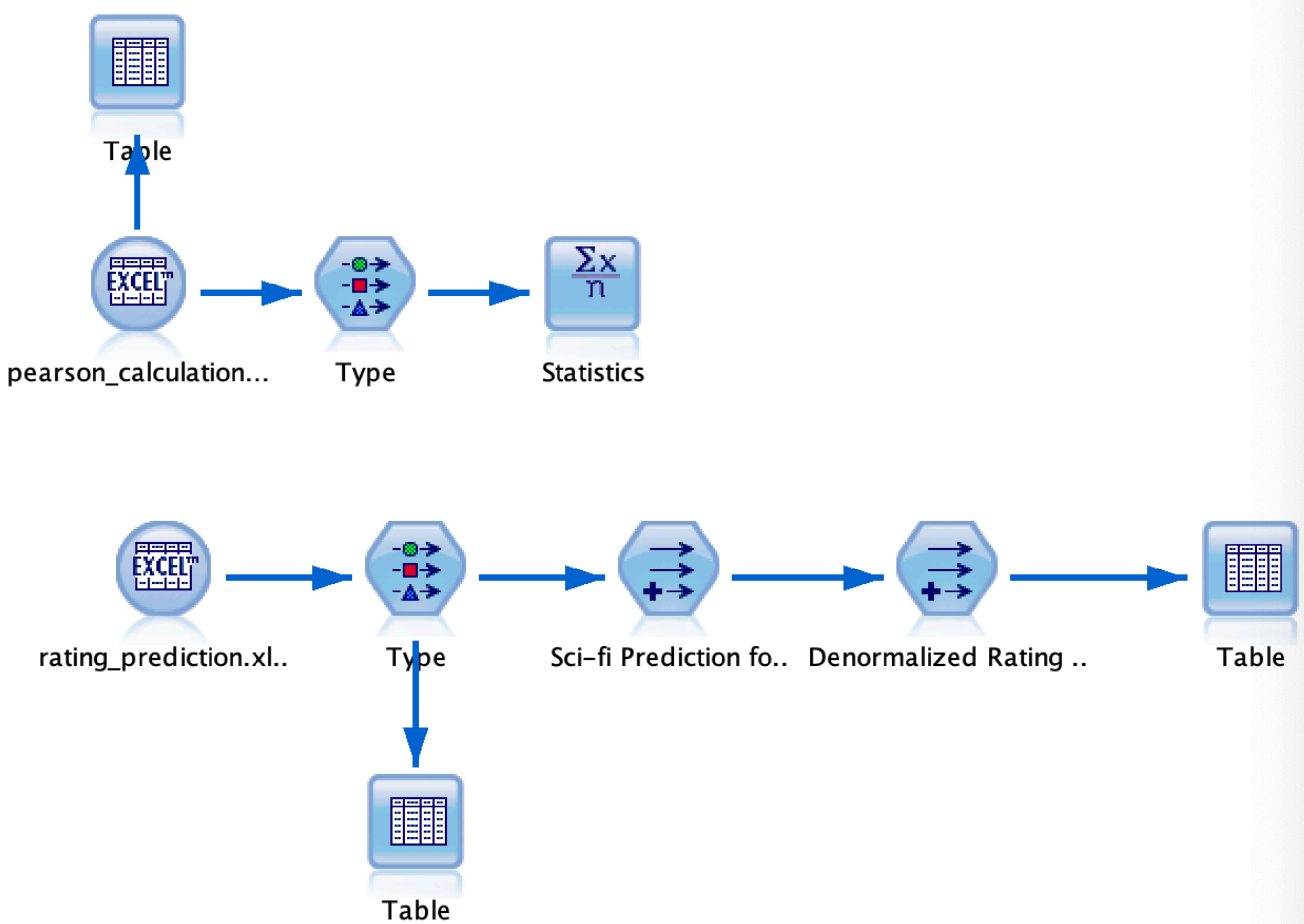
Help people discover new genres

Recommend a movie genre that a user normally doesn’t watch but might like, by identifying other users with similar tastes using the Pearson correlation

Concept



Preparation



34.000000

Statistics

Count	12
Mean	-0.000
Min	-2.651
Max	1.349
Range	4.000
Variance	1.765
Standard Deviation	1.328
Standard Error of Mean	0.383

Pearson Correlations

2.000000	0.843	Strong
3.000000	0.785	Strong
4.000000	-0.324	Weak
5.000000	0.320	Weak
6.000000	0.509	Medium
7.000000	0.073	Weak
8.000000	0.385	Medium
9.000000	-0.251	Weak
10.000000	0.381	Medium
11.000000	0.598	Medium
12.000000	0.344	Medium
13.000000	0.600	Medium
14.000000	0.657	Medium
15.000000	0.742	Strong
16.000000	-0.338	Medium
17.000000	-0.208	Weak
18.000000	0.354	Medium
19.000000	-0.443	Medium
20.000000	0.286	Weak
21.000000	0.246	Weak
22.000000	0.178	Weak

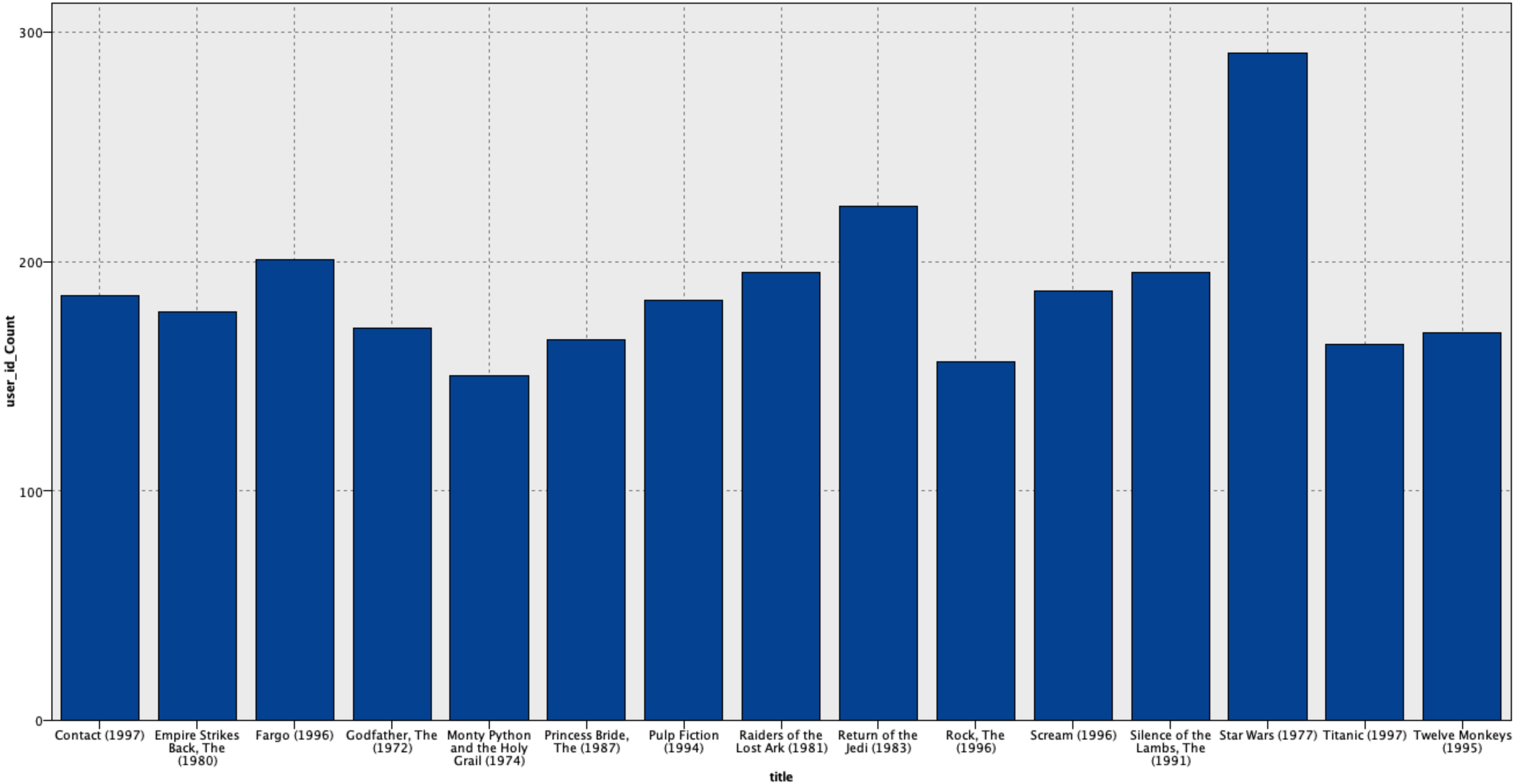
Results

	user_id	correlation	Sci–fi_rating	Sci–fi Prediction for user_id 34	Denormalized Rating for user_id 34
1	162	0.939	–0.046	0.171	3.822
2	74	0.917	–0.001	0.171	3.822
3	438	0.913	0.550	0.171	3.822
4	113	0.911	0.557	0.171	3.822
5	674	0.909	–0.055	0.171	3.822
6	634	0.904	0.321	0.171	3.822
7	772	0.902	0.308	0.171	3.822
8	117	0.897	0.169	0.171	3.822
9	76	0.893	0.110	0.171	3.822
10	566	0.893	–0.199	0.171	3.822

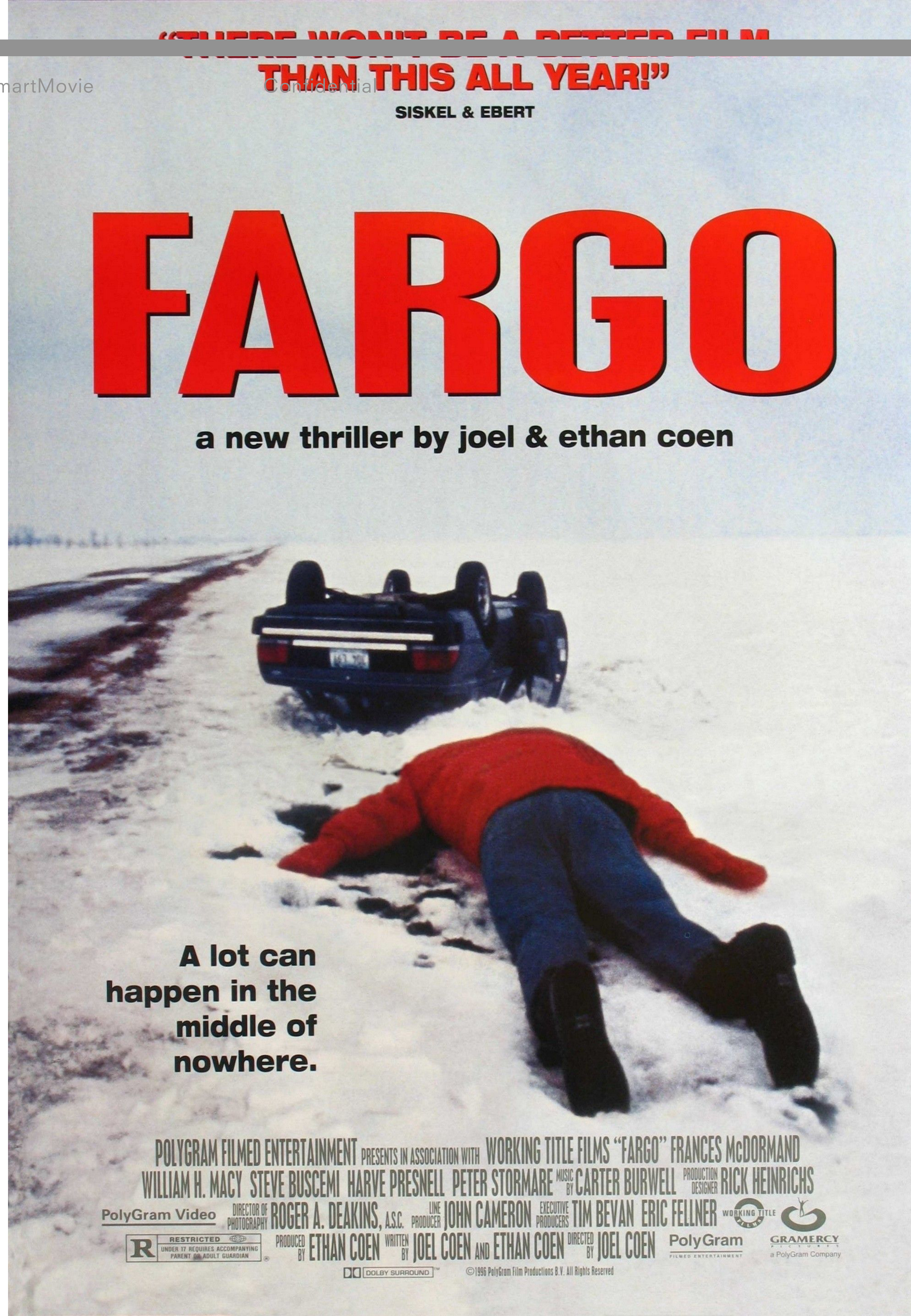
We think you'll like Sci-fi, even though you never rated it!

Movie based theme park rides

Popular movies among age 19-40




UNIVERSAL


THAN THIS ALL YEAR!”
SISKEL & EBERT
FARGO
a new thriller by joel & ethan coen
A lot can happen in the middle of nowhere.
POLYGRAM FILMED ENTERTAINMENT PRESENTS IN ASSOCIATION WITH WORKING TITLE FILMS "FARGO" FRANCES McDORMAND
WILLIAM H. MACY STEVE BUSCEMI HARVE PRESNELL PETER STORMARE MUSIC BY CARTER BURWELL PRODUCTION DESIGNER RICK HEINRICHS
PolyGram Video
DIRECTOR OF PHOTOGRAPHY ROGER A. DEAKINS, A.S.C. LINE PRODUCER JOHN CAMERON EXECUTIVE PRODUCERS TIM BEVAN ERIC FELLNER
PRODUCED BY ETHAN COEN WRITTEN BY JOEL COEN AND ETHAN COEN DIRECTED BY JOEL COEN
R RESTRICTED UNDER 17 REQUIRES ACCOMPANYING PARENT OR ADULT GUARDIAN
PolyGram
GRAMERCY
DOLBY SURROUND
©1996 PolyGram Film Productions B.V. All Rights Reserved

Business goal 05

Recommend new theme park rides by determining most popular movie for a selected age group

Find out the most popular movie for certain age group and create recommendations for the Theme park for making new attractive rides. Help in attracting a different age group to these theme parks using the dataset.

Conclusion

With only a little data, we helped you...

- Identify and create movies with most popular genres for target audiences
- Optimize marketing opportunities for target user groups
- Help users discover new movie genres
- Identify popular theme park ride ideas
- **Increase your profits!**

Hire us, thank you!

FAB 4 LLC

Appendices

Business Goals

1. Suggest genres for new movies for targeted age groups
2. Determine unique movie genre combinations to produce
3. Provide marketing opportunities by reaching the most active users in their workplace
4. Help people discover new genres that they haven't seen before
5. Recommend new theme park rides by determining most popular movie for a selected age group

Success Criteria

1. Identify most popular genre of movie for the most popular age group

2. Identify most popular movie genre combinations and how many movies are already produced

3. Display most active users within the most active workplaces

4. Predict the genre rating for a particular user by using pearson correlation and identifying their top 10 nearest neighbors

5. Identify most popular movie for a certain age group and create recommendations for theme park rides.

Data Analysis

Main Data Set

Data Analysis - Main Dataset

- This table contains:
    - User ID
    - Movie ID
    - Rating
    - Title
    - Genre
    - Age
    - Gender
    - Occupation

Data Analysis - Main Dataset

- This data can be used to create user profiles. This data also links movie id’s to movie titles. Moreover, the genres for the movies are given in this dataset

- All data is important

Data Analysis

User Average Rating per Genre

Data Analysis - User Average Rating per Genre

- This table contains:
  - User id
  - Average per genre
- This data can be used to find out which user has rated which genres higher than others. The data needs to be normalised.
- All data is important

Data Analysis

User Movie Genre

Data Analysis - User Movie Genre

- This table contains:
  - User ID
  - Movie ID
  - Genre
- Since the main dataset includes genres per movie ID, this dataset does not need to be used within this analysis.

Data Preparation

Main Dataset

Data Preparation - Main Dataset

- Imported file into SPSS Modeller
- Looked at the data and changed types:
    - Occupation to nominal
    - Gender to flag
    - Genres to flag

Field	Measurement	Values	Missing	Check	Role
user_id	Continuous	[2,943]		None	Input
movie_id	Continuous	[2,1682]		None	Input
rating	Continuous	[1,5]		None	Input
title	Typeless			None	None
Unkown	Flag	1/0		None	Input
Action	Flag	1/0		None	Input
Adventure	Flag	1/0		None	Input
Animation	Flag	1/0		None	Input
Childrens	Flag	1/0		None	Input
Comedy	Flag	1/0		None	Input
Crime	Flag	1/0		None	Input
Documentary	Flag	1/0		None	Input
Drama	Flag	1/0		None	Input
Fantasy	Flag	1/0		None	Input
Film-Noir	Flag	1/0		None	Input
Horror	Flag	1/0		None	Input
Musical	Flag	1/0		None	Input
Mystery	Flag	1/0		None	Input
Romance	Flag	1/0		None	Input
Sci-Fi	Flag	1/0		None	Input
Thriller	Flag	1/0		None	Input
War	Flag	1/0		None	Input
Western	Flag	1/0		None	Input
age	Continuous	[7,73]		None	Input
gender	Flag	M/F		None	Input
occupation	Nominal	administrator,artist,d...		None	Input

- Note: Title is typeless because there are too many different ones. This does not influence the dataset and was, therefore, kept like typeless.

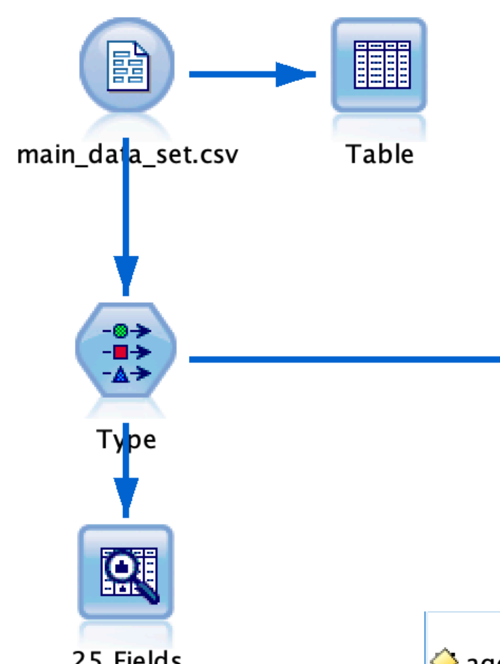
Maaike

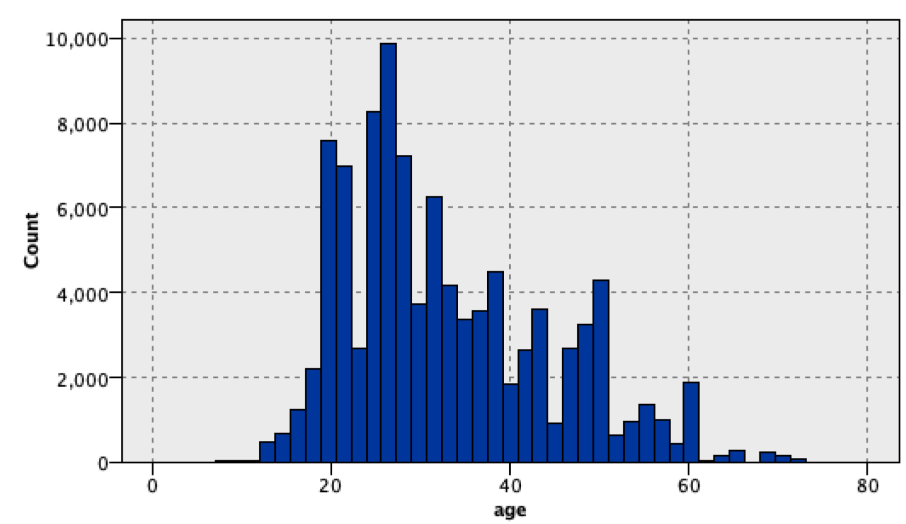
Data Preparation - Main Dataset

- Analysed data: only age could have outliers. Looking at the age graph below, we can determine that we would like to keep the parabola, which is a good representation of the population in general. Through the audit node, the following measurements were taken in regards to age:

- Mean : 32.999

- Std: 11.573





age		Continuous	7	73	32.999	11.573

Maaike

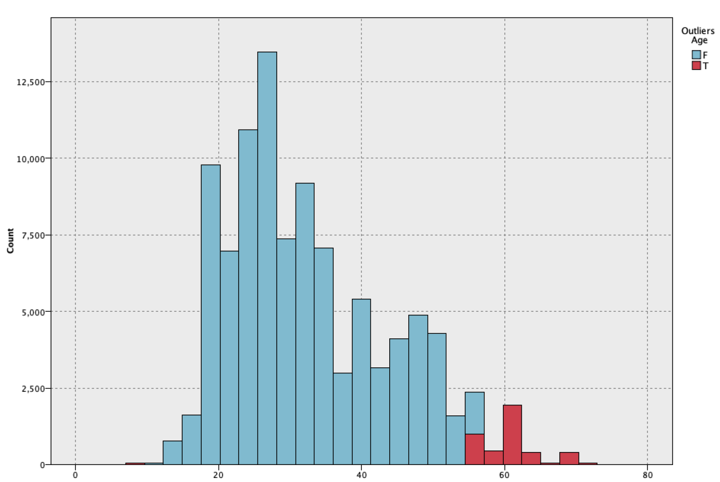
Data Preparation - Main Dataset

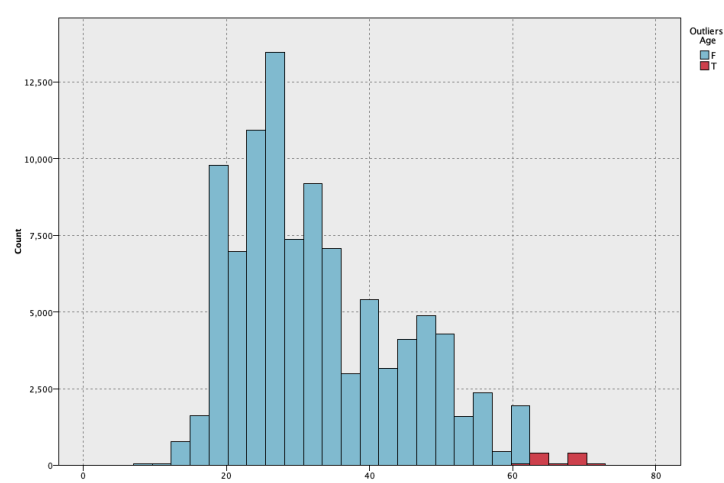
- Several equations were tried in order to determine which one would exclude the outliers:

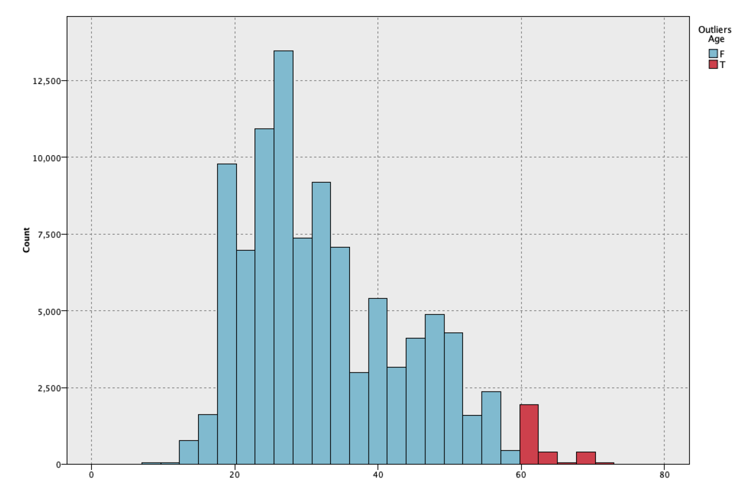
  - 1: ’age'<(32.999-1.5*11.573) or ‘age’>(32.999+1.5*11.573)

  - 2: ’age'<(32.999-1.5*11.573) or ‘age’>(32.999+1.5*11.573)

  - 3: ’age'<(32.999-1.5*11.573) or ‘age'>(32.999+1.5*11.573)







Maaike

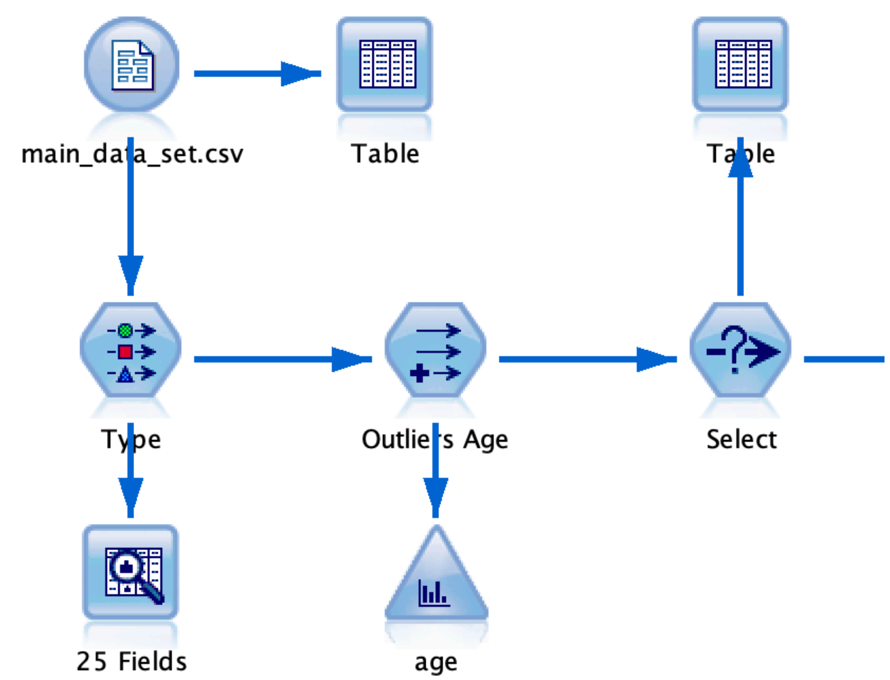
Data Preparation - Main Dataset

- Outliers were detected:

Value	Proportion	%	Count
F		97.17	96462
T		2.83	2814

- Through the select node, the outliers were deleted.

- The database went from 99,276 records to 96,462 records



Maaike

Data Preparation - Main Dataset

- Dataset was checked for empty ratings by checking if rating was equal to 0. Moreover, it was checked if all movies had at least one genre.

- This was done through the derive node by creating a flag for any entry that complies with the following equation:

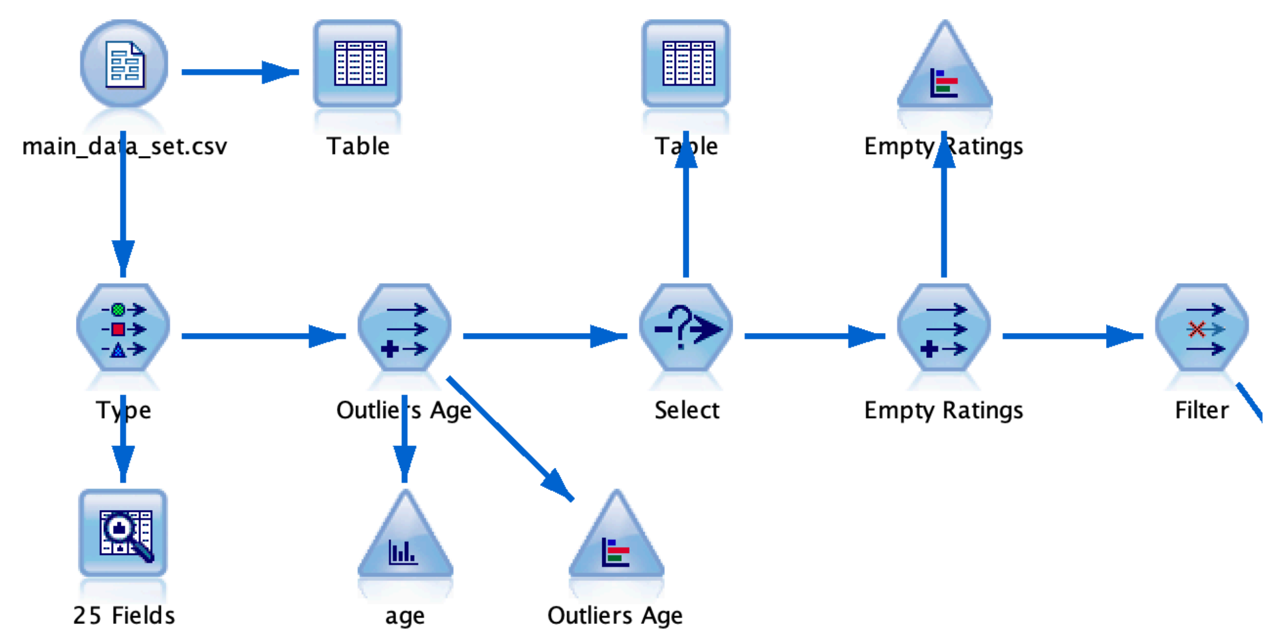
  - 'Unkown'=0 and 'Action'=0 and 'Adventure'=0 and 'Animation'=0 and 'Childrens'=0 and 'Comedy'=0 and 'Crime'=0 and 'Documentary'=0 and 'Drama'=0 and 'Fantasy'=0 and 'Film-Noir'=0 and 'Horror'=0 and 'Musical'=0 and 'Mystery'=0 and 'Romance'=0 and 'Sci-Fi'=0 and 'Thriller'=0 and 'War'=0 and 'Western'=0 or rating = 0

Data Preparation - Main Dataset

- After checking the flags, it turns out that all entries had a rating and every movie had at least one genre assigned to it.

- At the end, the record count was deleted from the database through the filter node

Value	Proportion	%	Count
F		100.0	96462



Maaike

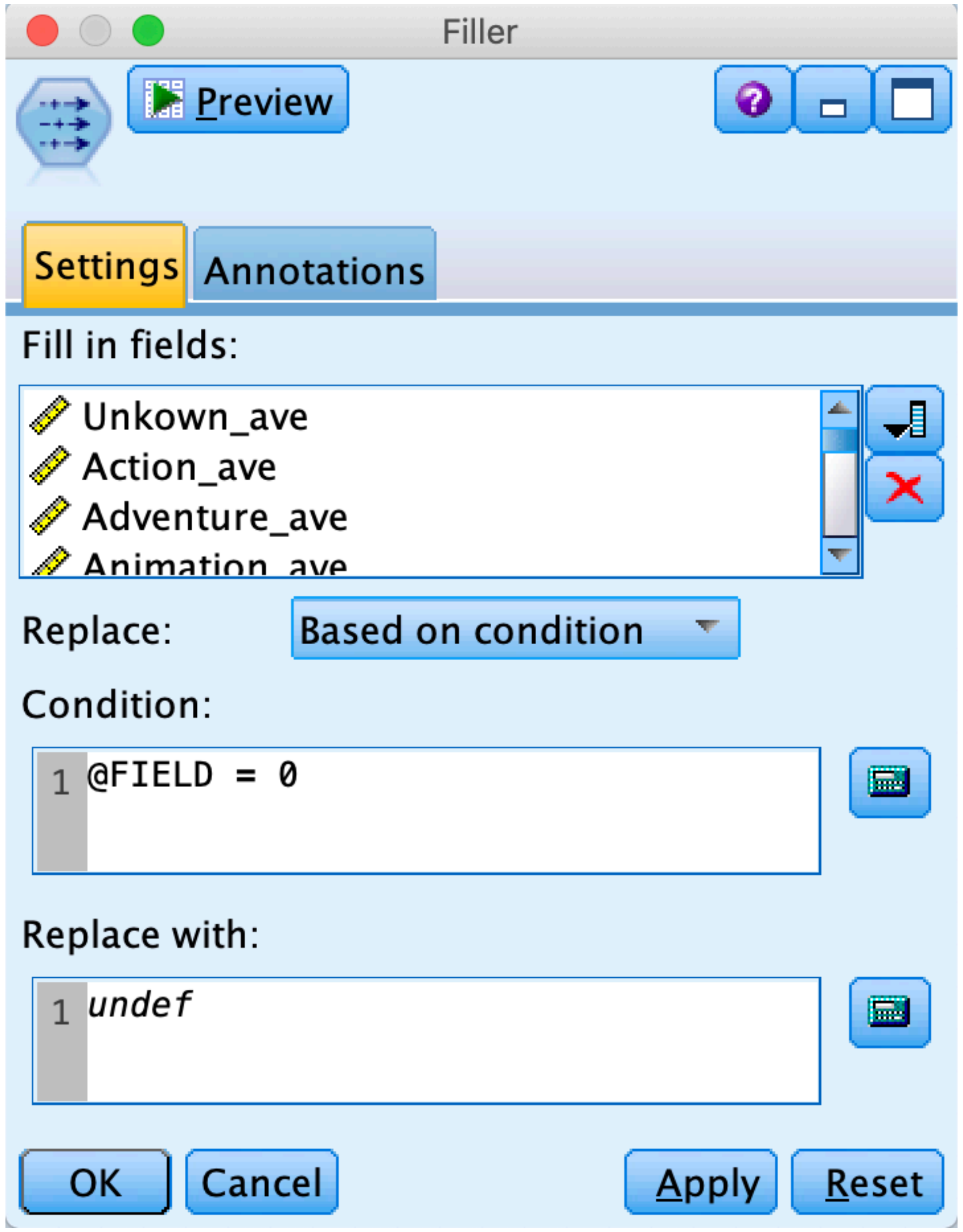
Data Preparation

User Average Rating per Genre

Data Preparation - User Average Rating per Genre

- Within the dataset, there are many users have a 0 average rating for a genre. This means that they have never rated a movie within this genre (since all ratings are between 1 and 5).

- In order to ensure that these zeros are not taken into account when doing calculations, the filler node is used to transform all zeroes to $null$ values for all genres



Maaike

Data Preparation - User Average Rating per Genre

- At the end, all data will need to be normalised. This is done because there can be a huge difference in ratings given depending on the type of person: a positive person might rate all 4’s and 5’s, whereas a more negative person can give ratings of 2’s and 3’s, but enjoy the movies the same.

- By normalising the data, you can see how the user’s rating is compared to their average rating. Above 0 is better than usual, below 0 is worse than usual.

- For this, we will need to know the average of the averages of the user. The user can love one genre and hate the other: to gain a better understanding of how the user rates, all averages need to be taken into account.

Data Preparation - User Average Rating per Genre

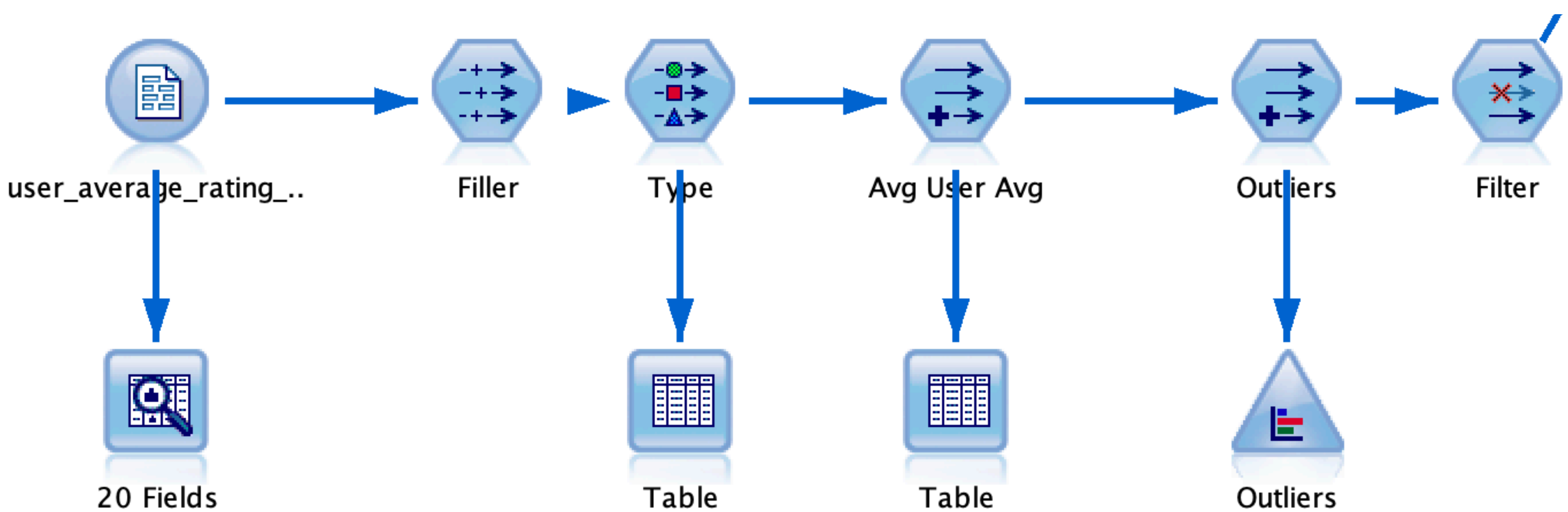
- A derive node is used to create a new field: avg user avg (average user average).
- First, all averages are added through the following formula:
  - (sum_n(@FIELDS_BETWEEN(Unkown_ave,Western_ave)))
- Then, this number needs to be divided by all the genres that have a value. Therefore, all null values need to be subtracted from the 19 genres. This is done through the following formula:
  - (19 - (count_nulls(@FIELDS_BETWEEN(Unkown_ave,Western_ave))))

Data Preparation - User Average Rating per Genre

- It was checked if average user average was ever equal to '0' (meaning the user did not rate any items). This was not the case:

Value	Proportion	%	Count
F		100.0	942

- Through a filter node, the outliers column was deleted.



Maaike

Data Preparation

Merged Dataset

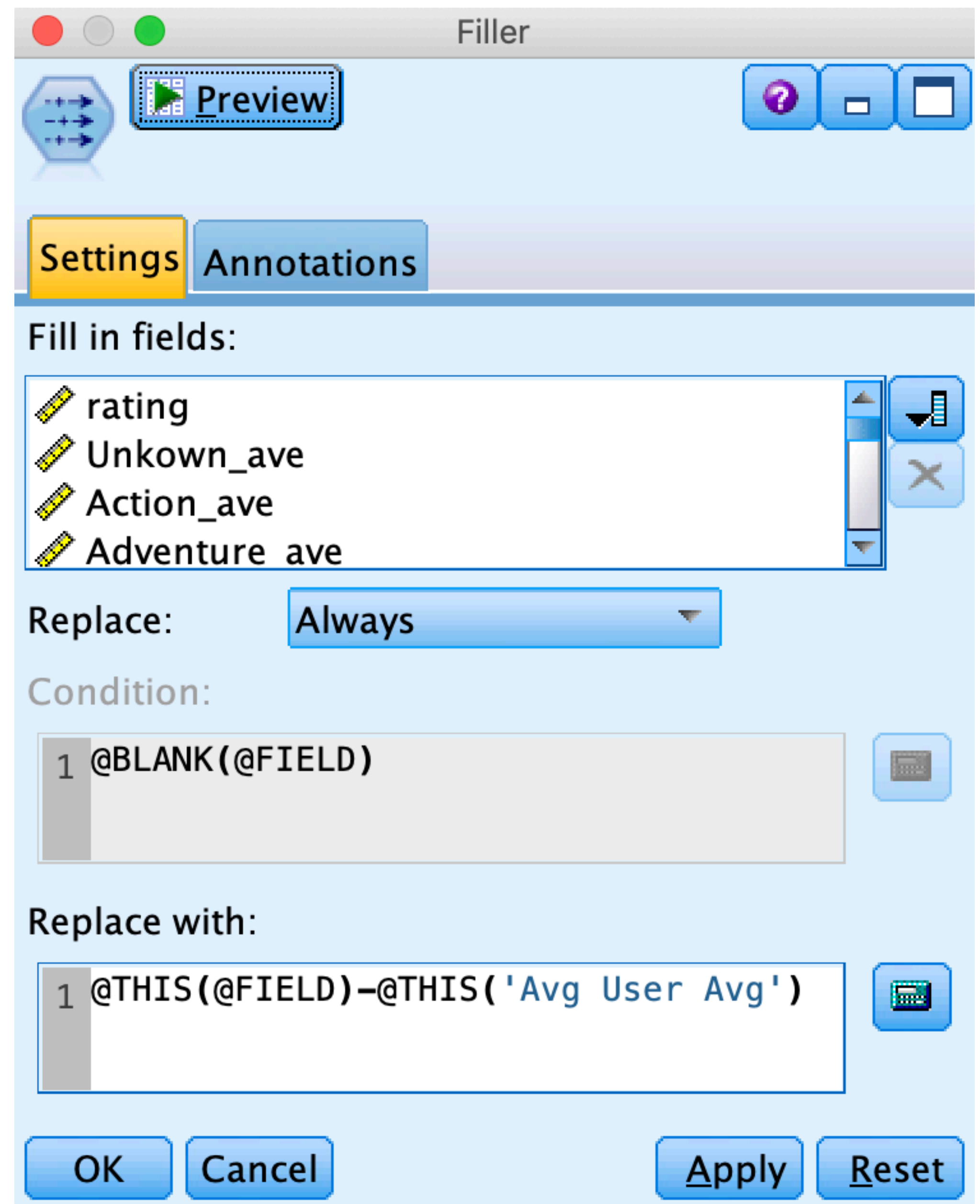
Data Preparation - Merged Dataset

- Both datasets were merged using the merge node. They were merged on the user-id.

- The dataset was sorted on the user-id (ascending order) through the ‘sort’ node.

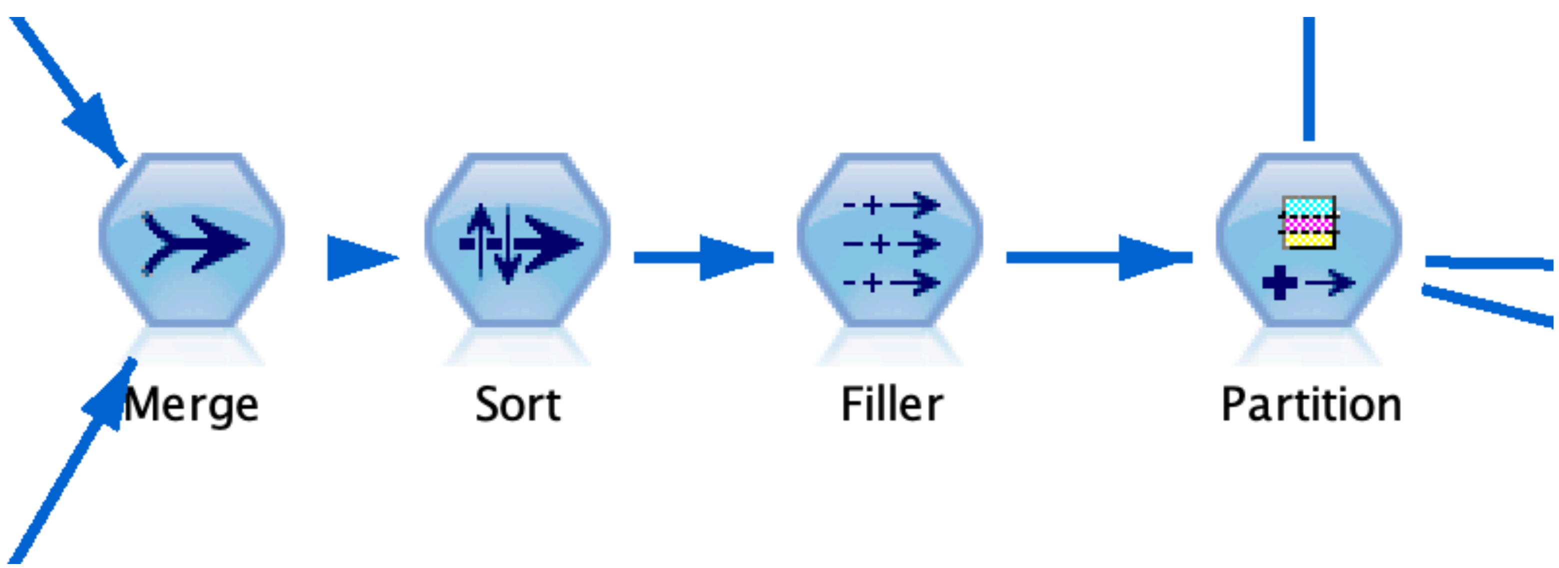
- Now, the dataset needs to be normalised. This can be done through the filler node. Here, the ratings and all the averages per user needed to be subtracted by the average user average:

@THIS(@FIELD)-@THIS('Avg User Avg’)



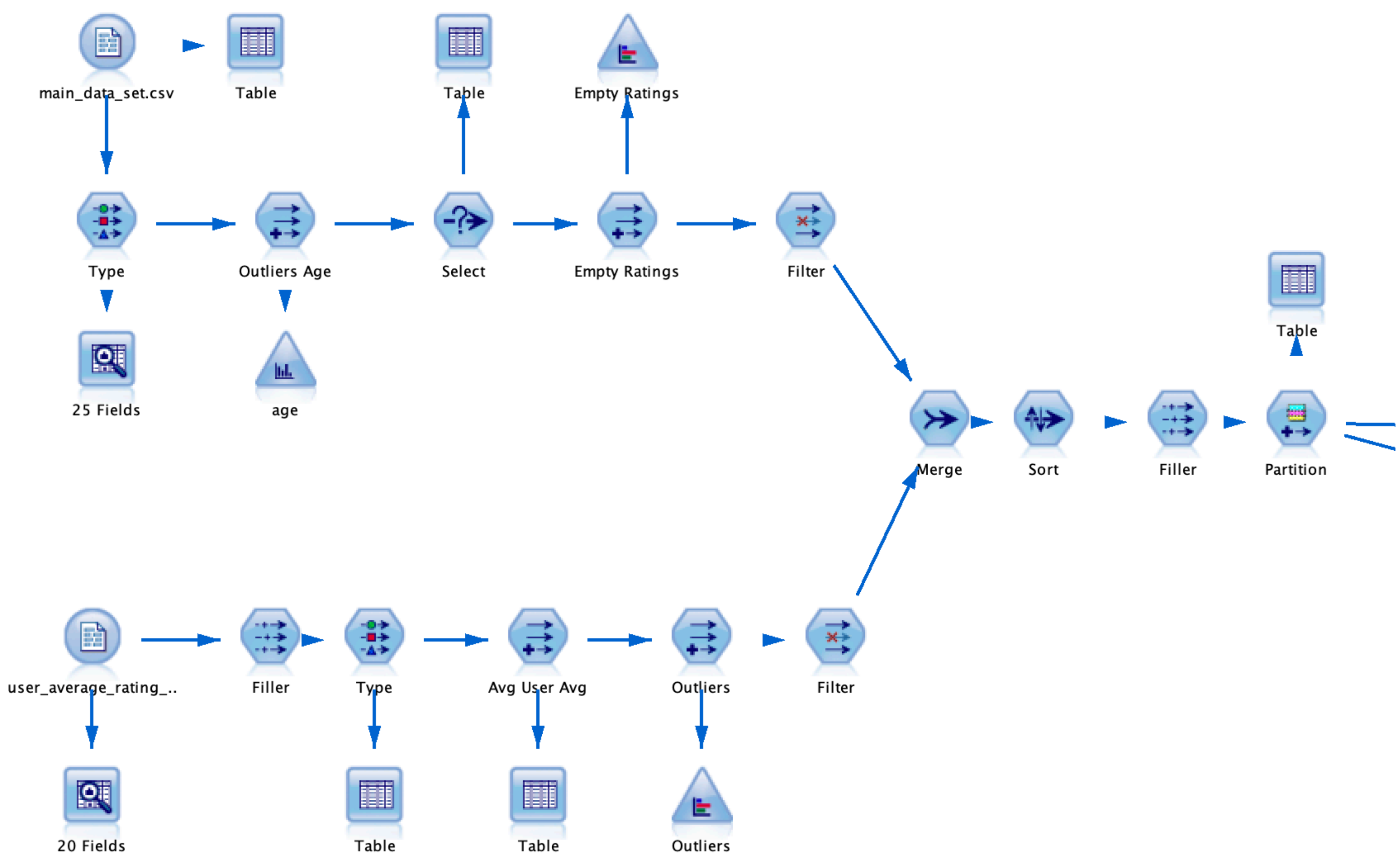
Data Preparation - Merged Dataset

- A partition node was added to create a testing set and a training set. The setup for the merged dataset was:



Data Preparation - Merged Dataset

- The setup for the entire data preparation process:



Maaike

Business Goal 1:
Determine the most popular genre for the most popular age group

Understand which is the most popular age group is and see which is their most popular genre of movie: based on this information, movie makers can produce popular genre movies for suitable audiences.

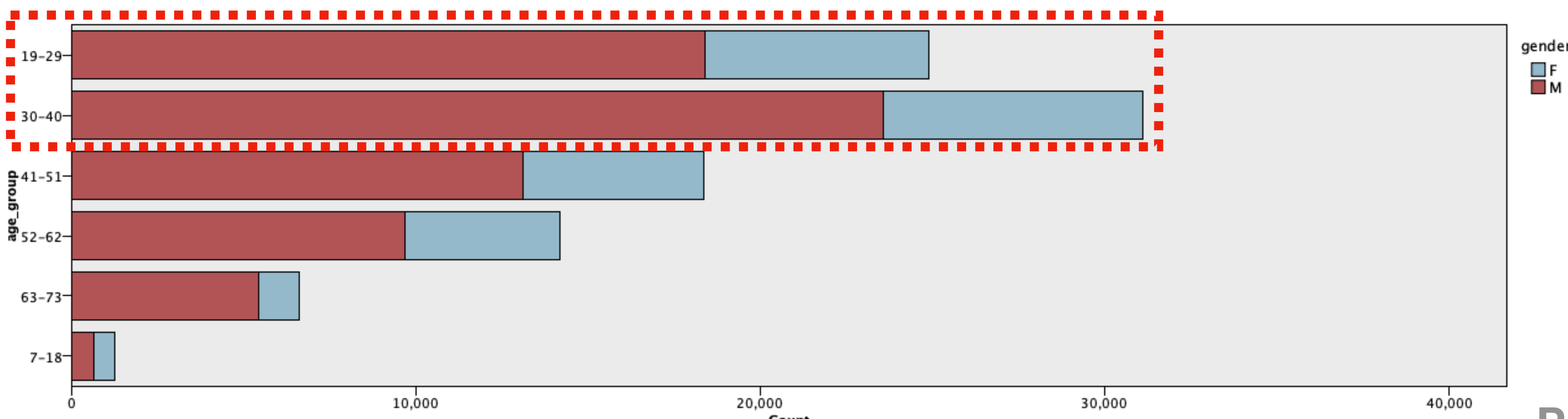
Business Goal 1: Determine the most popular genre for the most popular age group

- First, we must understand which is the most popular age group.

- Binned all the users ages into 6 groups.

Bin	Lower	Upper
1	>= 7	< 15.66666667
2	>= 15.66666667	< 24.33333333
3	>= 24.33333333	< 33
4	>= 33	< 41.66666667
5	>= 41.66666667	< 50.33333333
6	>= 50.33333333	<= 59

- Found out which was the most popular age group.

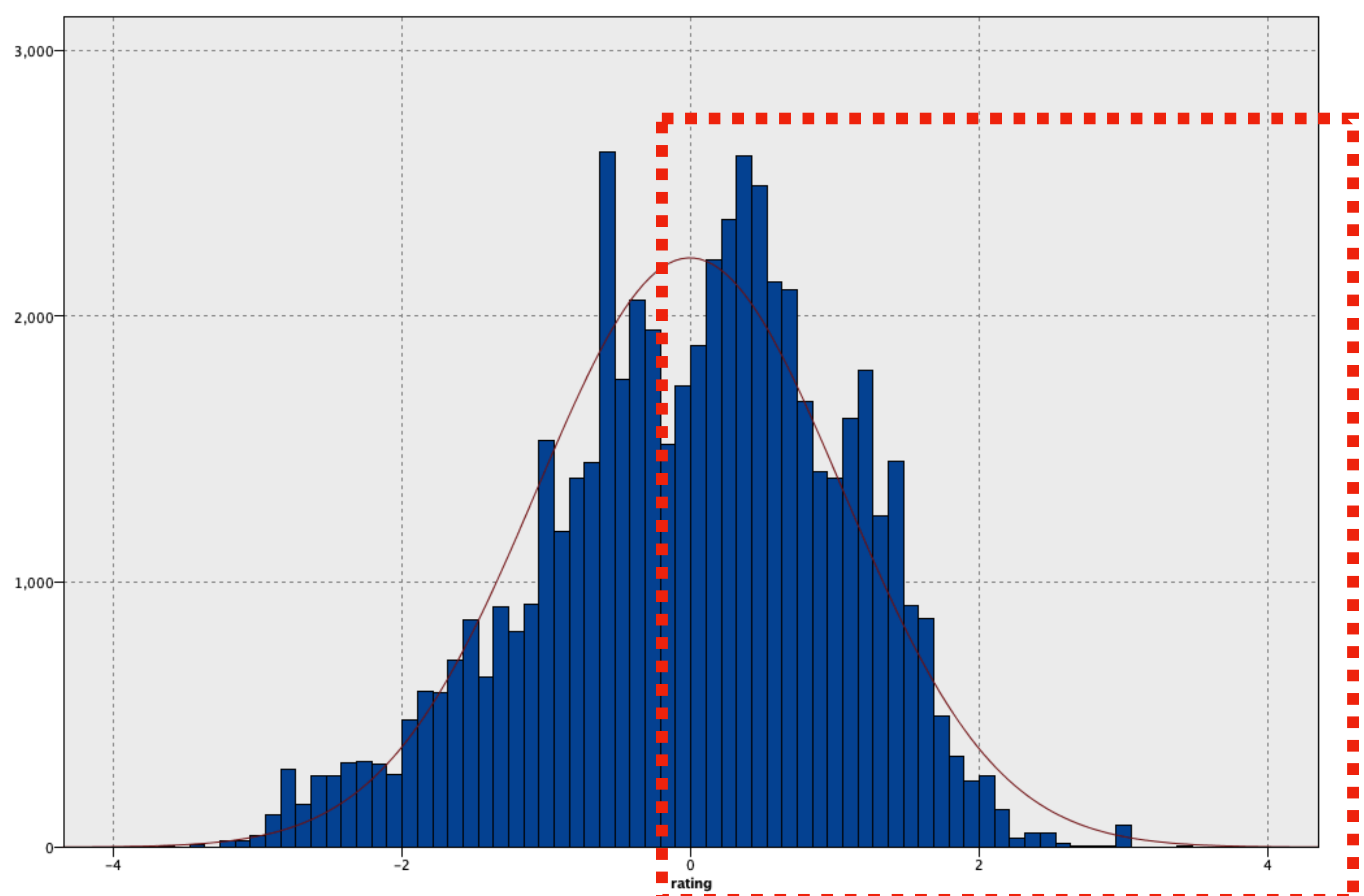


Pammu

Business Goal 1:
Determine the most popular genre for the most popular age group

- After looking at the graph, we decided to chose the popular age group of 19-40 years for a wider analysis.

- After analysing the graph, we decided to chose genres that have ratings more than 0.



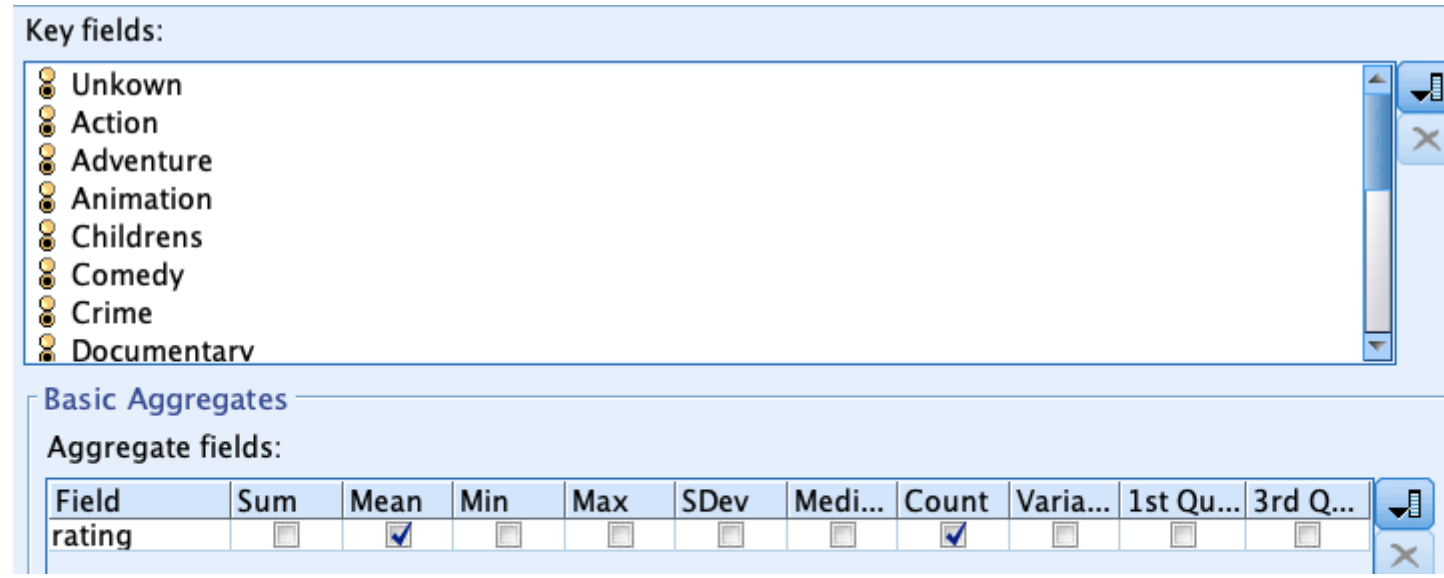

```
rating >= 0
```

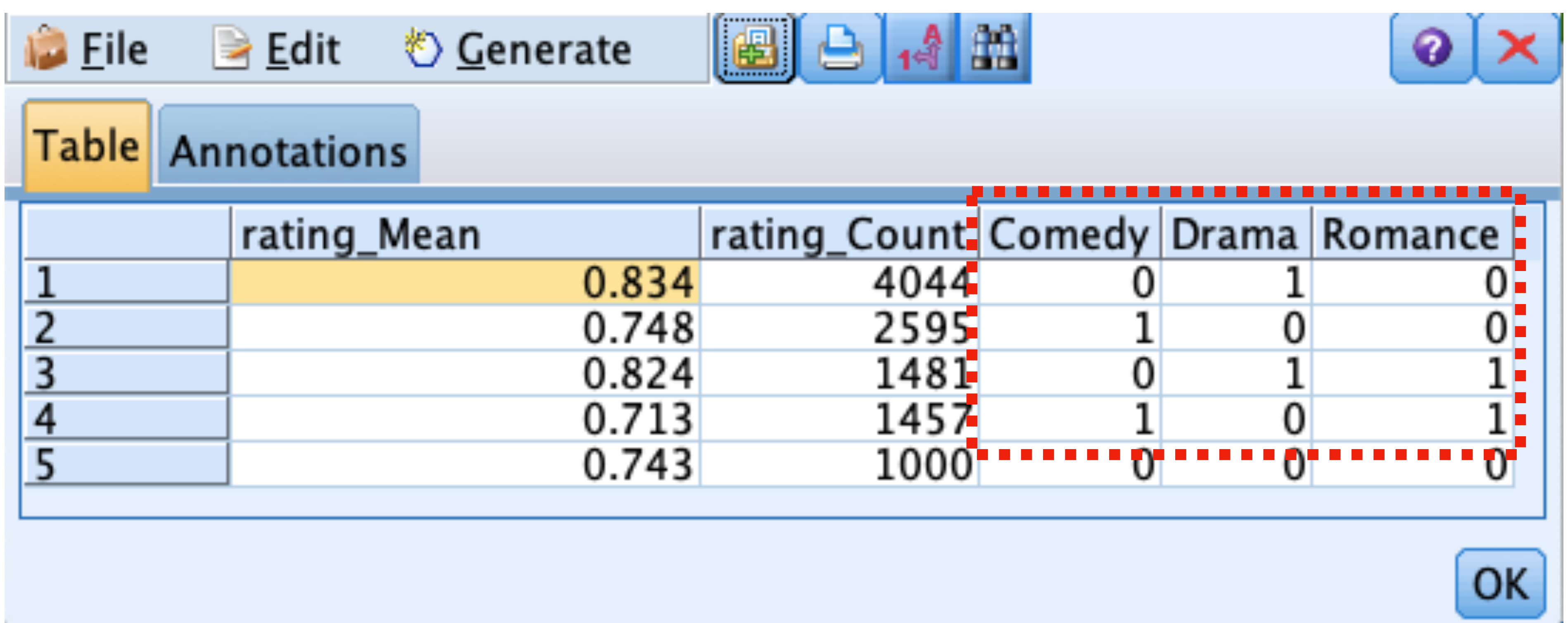
Pammu

Business Goal 1:
Determine the most popular genre for the most popular age group

- After that we checked the count of ratings given for each genres.

- After sorting through the table, and filtering it, we found a table which shows which are the 4 most popular genres for the selected age group.

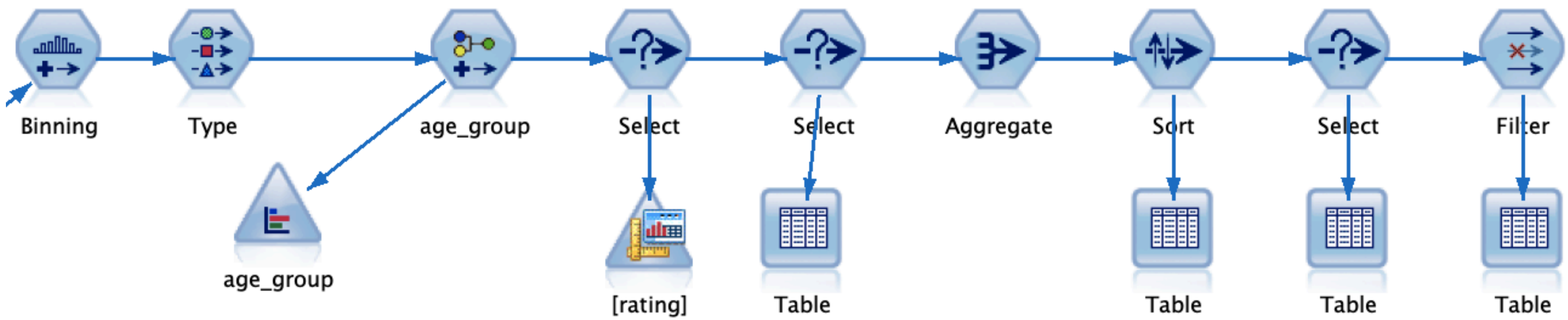




	rating_Mean	rating_Count	Comedy	Drama	Romance
1	0.834	4044	0	1	0
2	0.748	2595	1	0	0
3	0.824	1481	0	1	1
4	0.713	1457	1	0	1
5	0.743	1000	0	0	0

Pammu

Business Goal 1:
Determine the most popular genre for the most popular age group

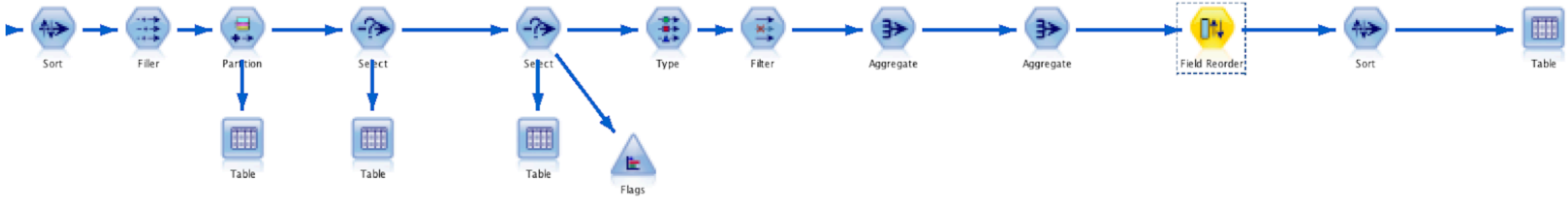


- The final setup for business goal 1 can be seen above. Of course, the data preparation will be added before this string in order to provide the complete result.

Business goal 2: Show which combinations of movie genres work well and show how many movies there are already produced in that genre

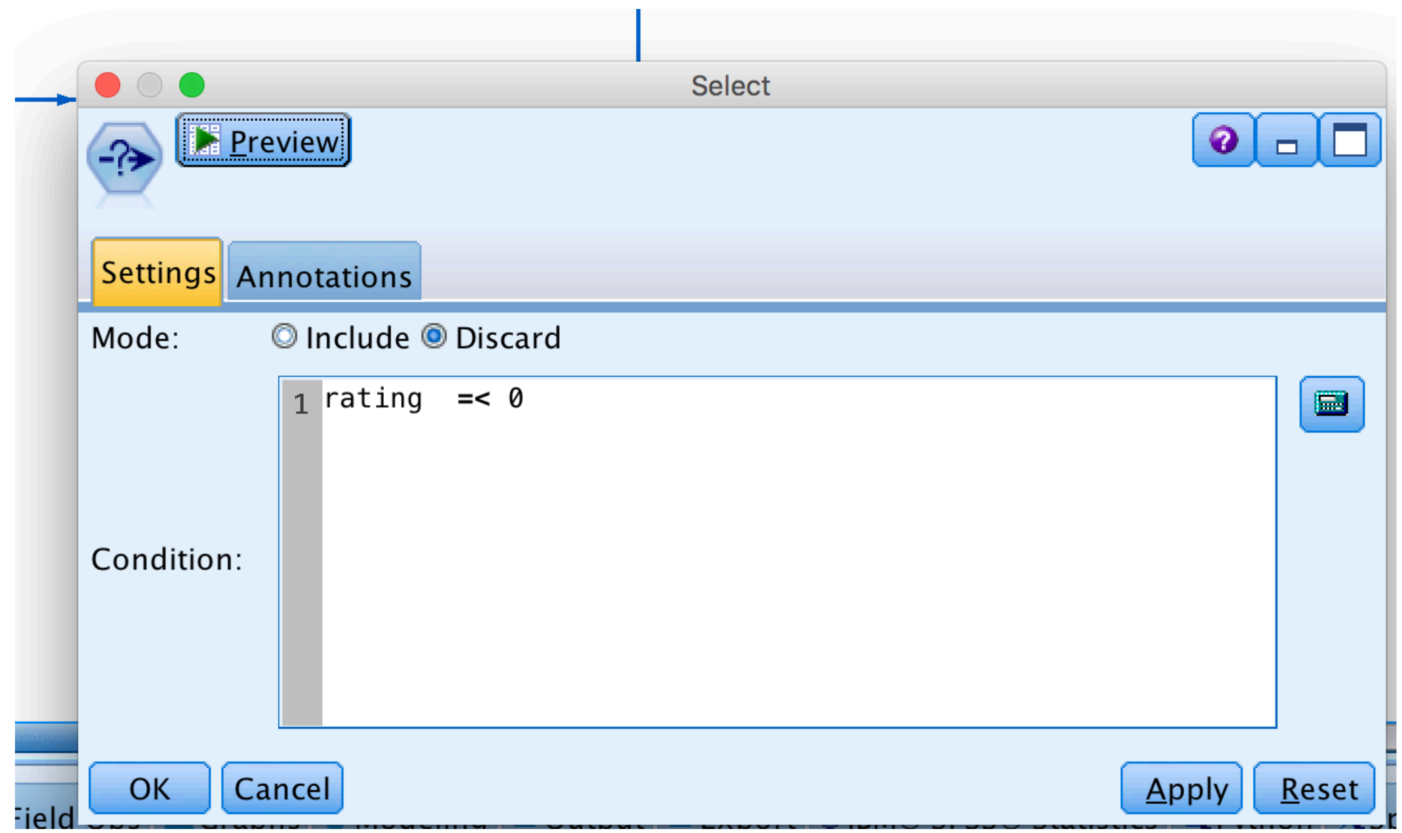
Understand which genres are the most popular movie genre combination and how many movies are produced, such that movie producer has quick insight and could determine in which genres combination could be option for a new movie.

Carmen

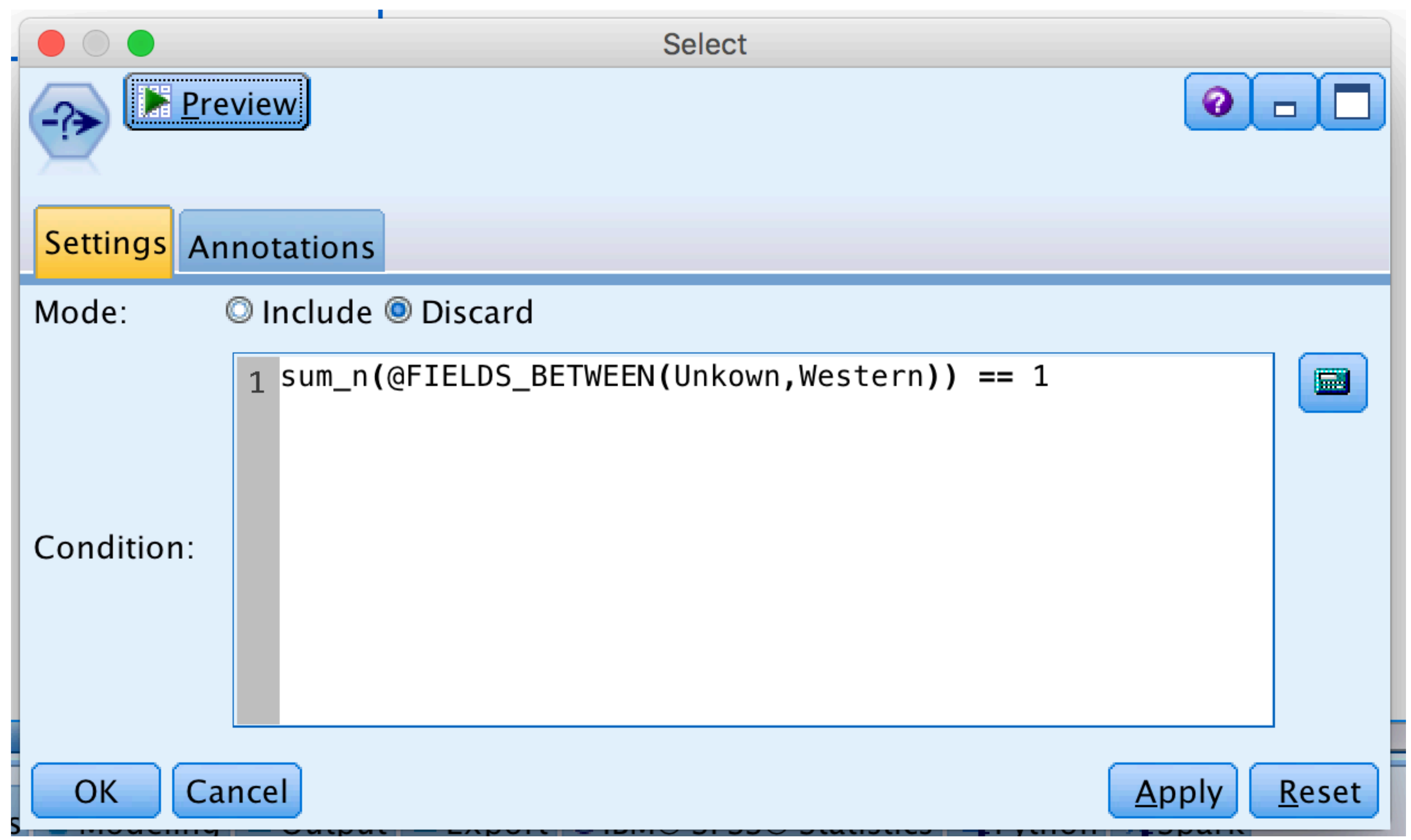

Sort
Filler
Partition
Select
Select
Type
Filter
Aggregate
Aggregate
Field Reorder
Sort
Table
Table
Table
Table
Flags

Carmen

- First we must understand what the popular movies are, therefore we need to select and discard(delete) all the not popular movies. According to the normalized values, every movie with a normalized rating above 0 is a popular movie.

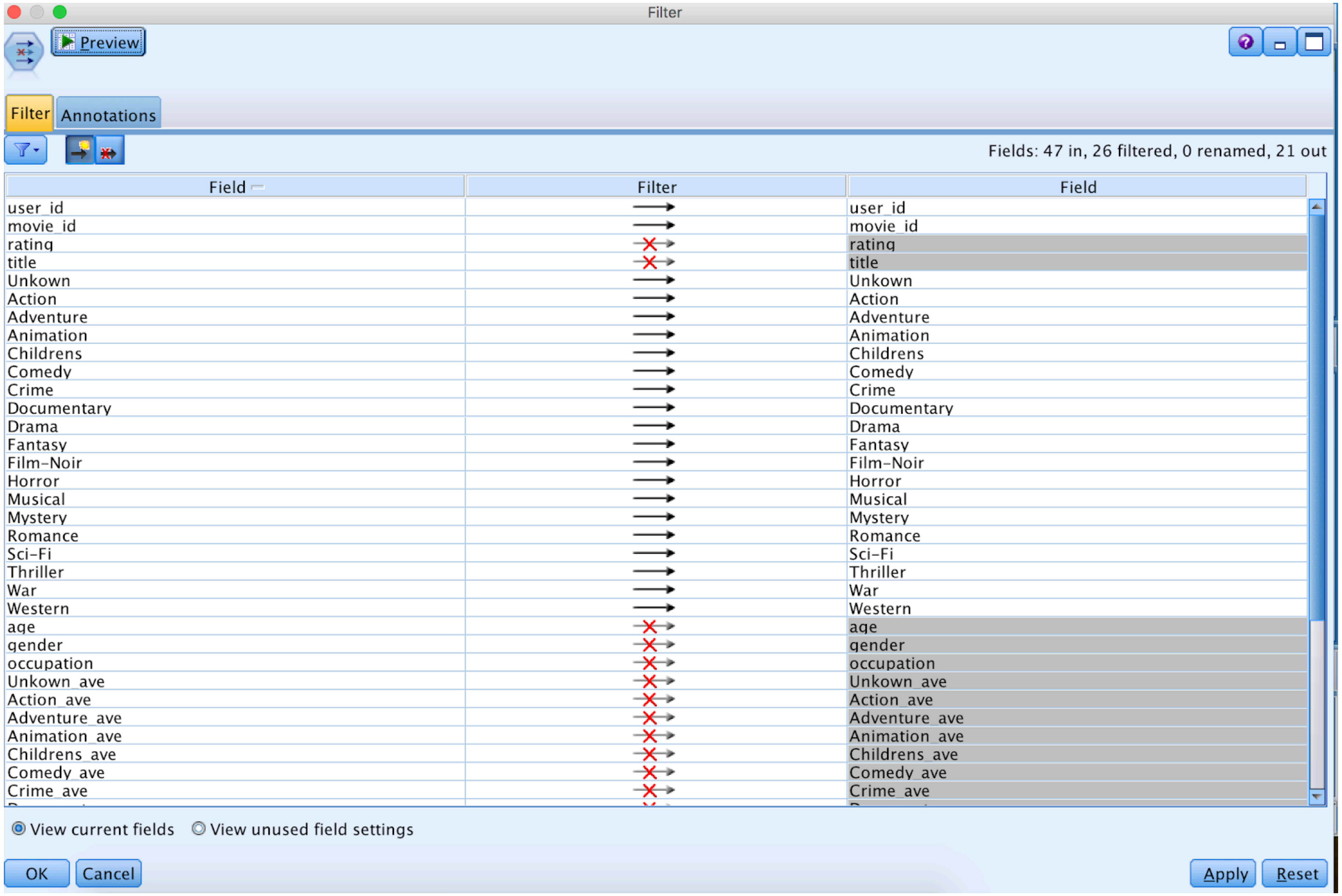


- After that we can start to determine the popular combinations. A combination always exist out of two componenter, therefore we can discard all user that only rated 1 genre. We can calculate that by taking the sum of all movies genres. If the sum adds up to 1, we know only one genre is used instead of two and then we can discard them.



Carmen

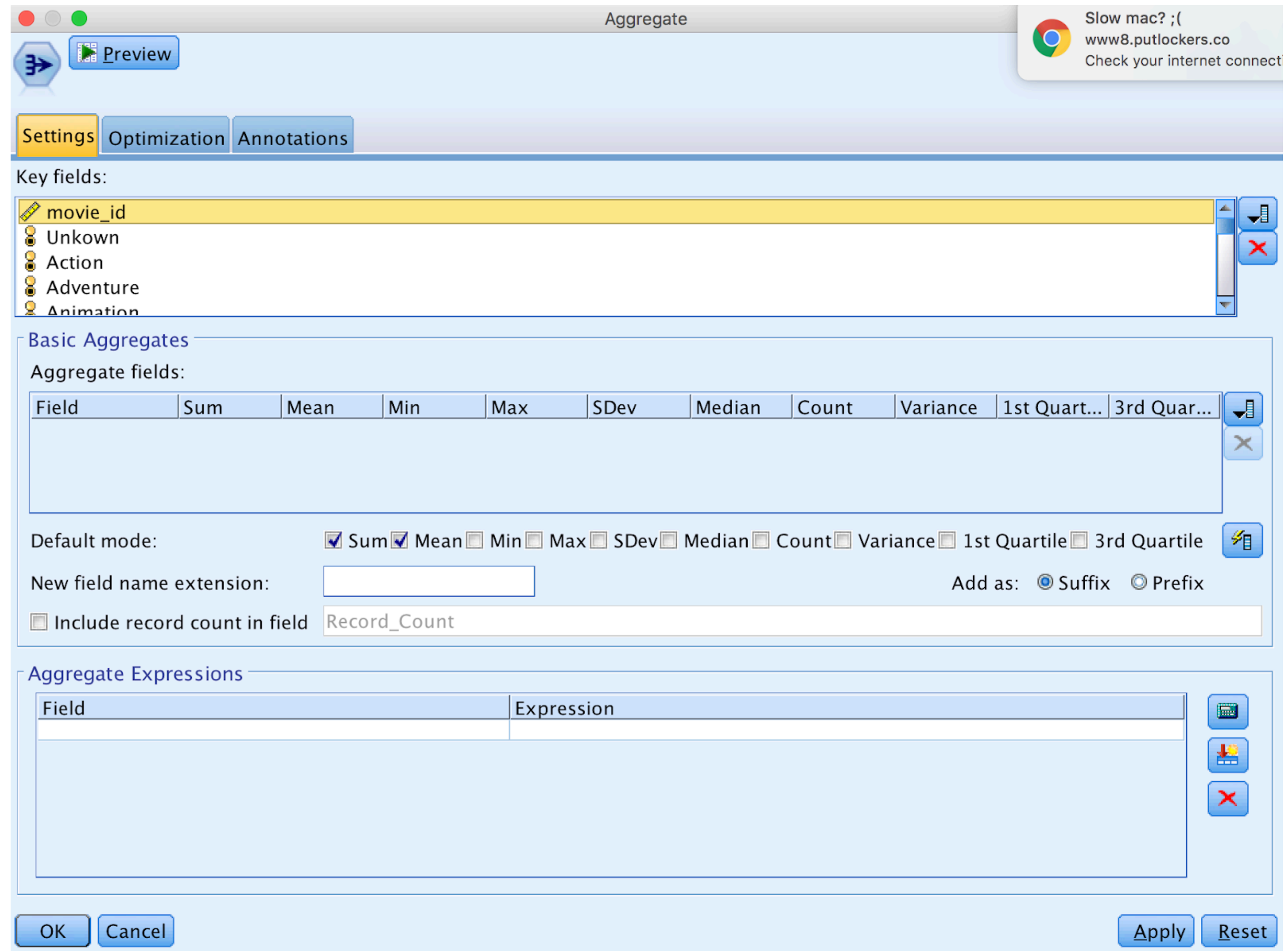
- After that we filter out all the stream we don't need, for a clearer result in the table

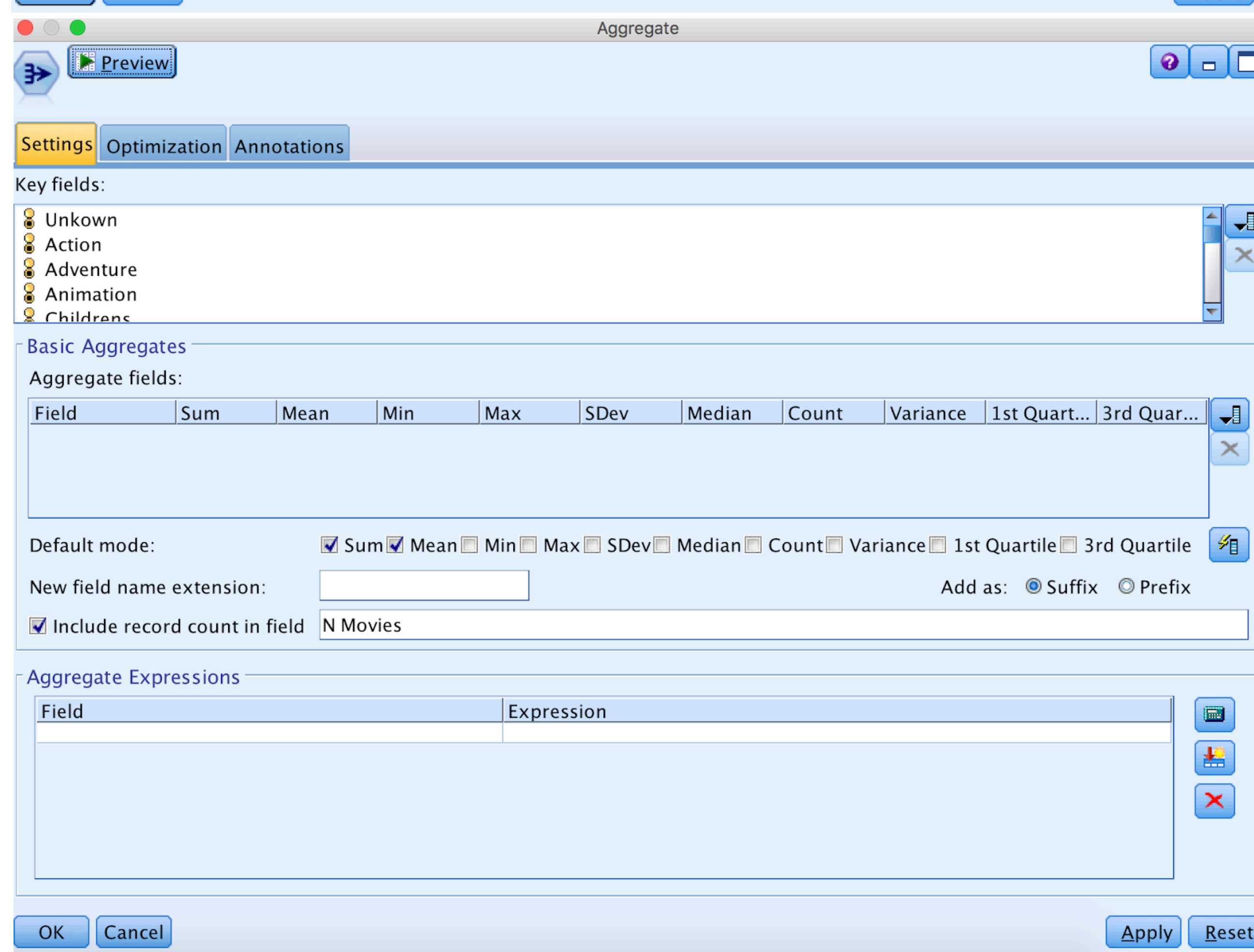


Carmen

After that we first aggregate on unique (popular) movies and the genre we have in the datafile.

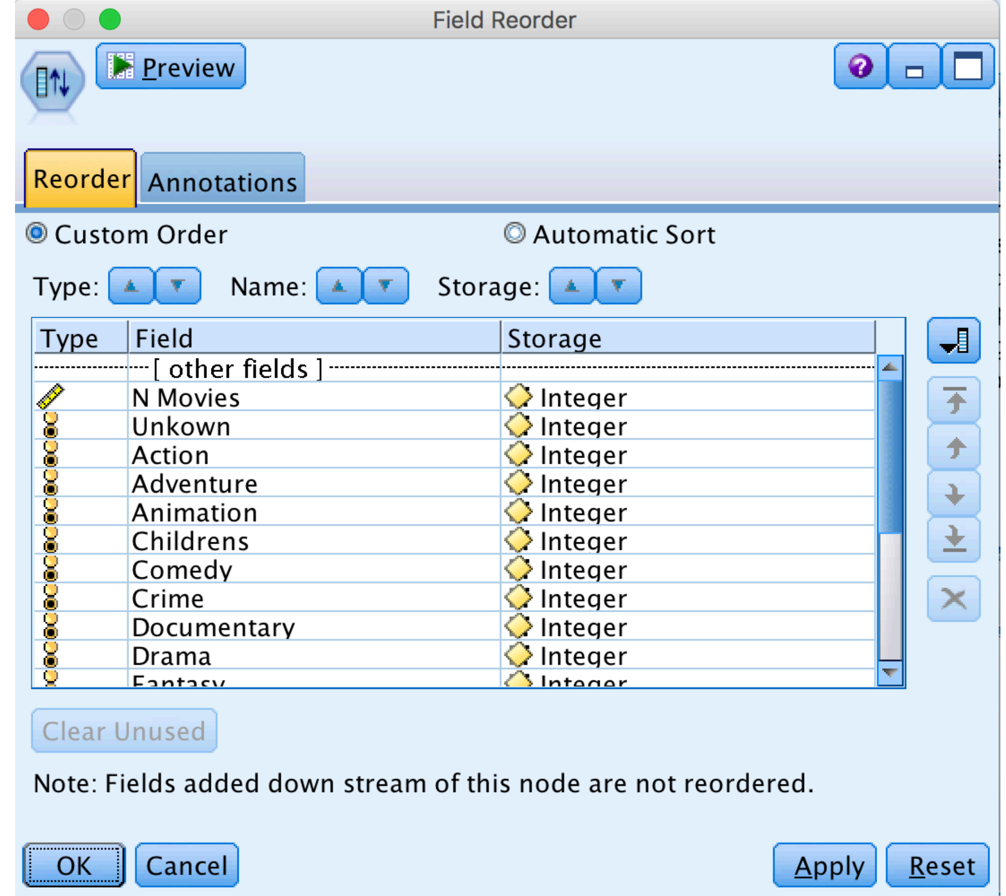
And then aggregate again on unique combinations of genres.

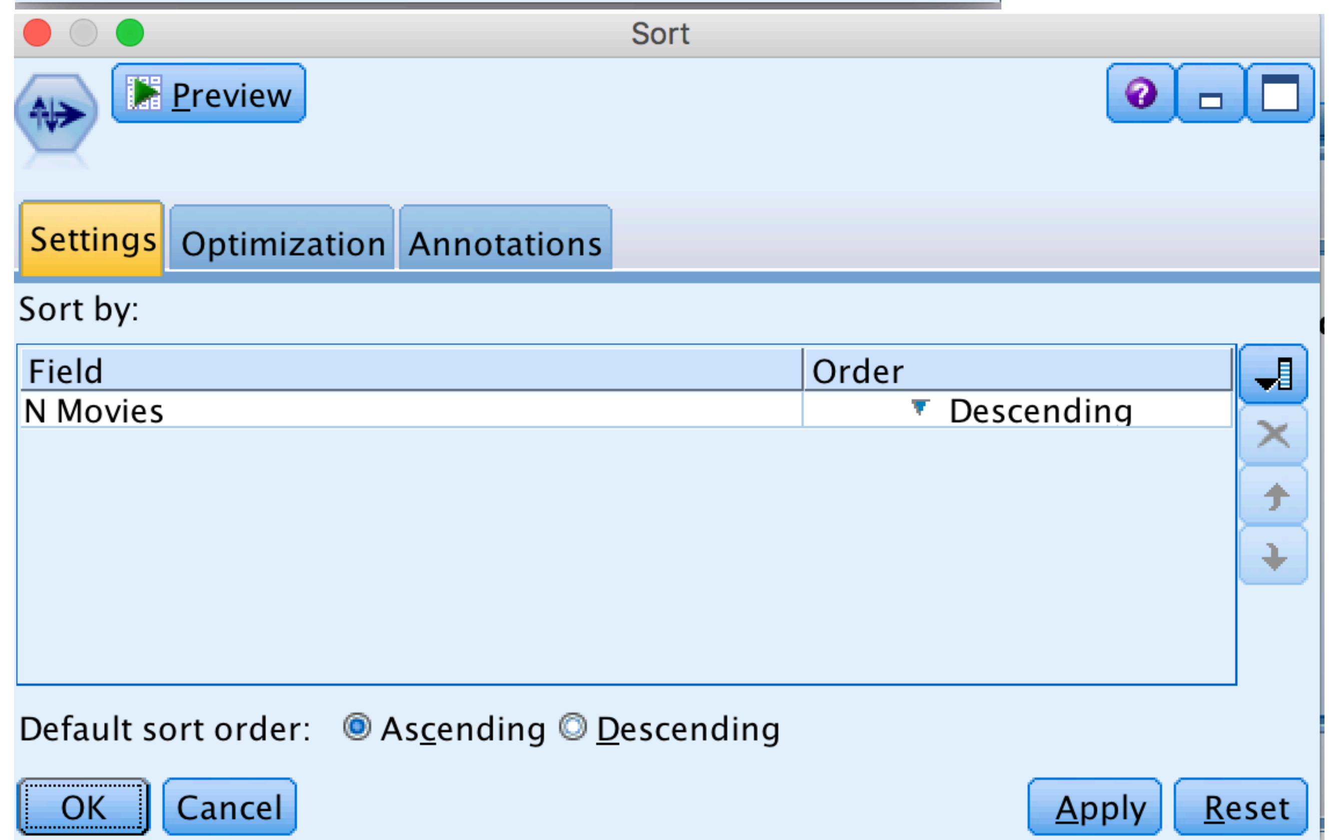




Carmen

After this the table should be made more insightful, therefore we change the order of the columns and order them from high to low.





Carmen

Table (20 fields, 197 records) #2

File Edit Generate

Table Annotations

	Unkown	Action	Adventure	Animation	Childrens	Comedy	Crime	Documentary	Drama	Fantasy	Film-Noir	Horror	Musical	Mystery	Romance	Sci-Fi	Thriller	War	Western	Record_Count
1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	2622
2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	2439
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1685
4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1377
5	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1317
6	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1278
7	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	877
8	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	849
9	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	834
10	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	829
11	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	762
12	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	761
13	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	682
14	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	649
15	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	634
16	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	593
17	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	590
18	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	555
19	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	549
20	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	509
21	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	463
22	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	396
23	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	380
24	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	359
25	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	355
26	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	349
27	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	318
28	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	310
29	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	296
30	0	1	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	1	0	282
31	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	264
32	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	261
33	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	254
34	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	241
35	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	238
36	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	235
37	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	225
38	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	217
39	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	216
40	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	215
41	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	207

OK

Carmen

Business Goal 3: Determine where the most active users work

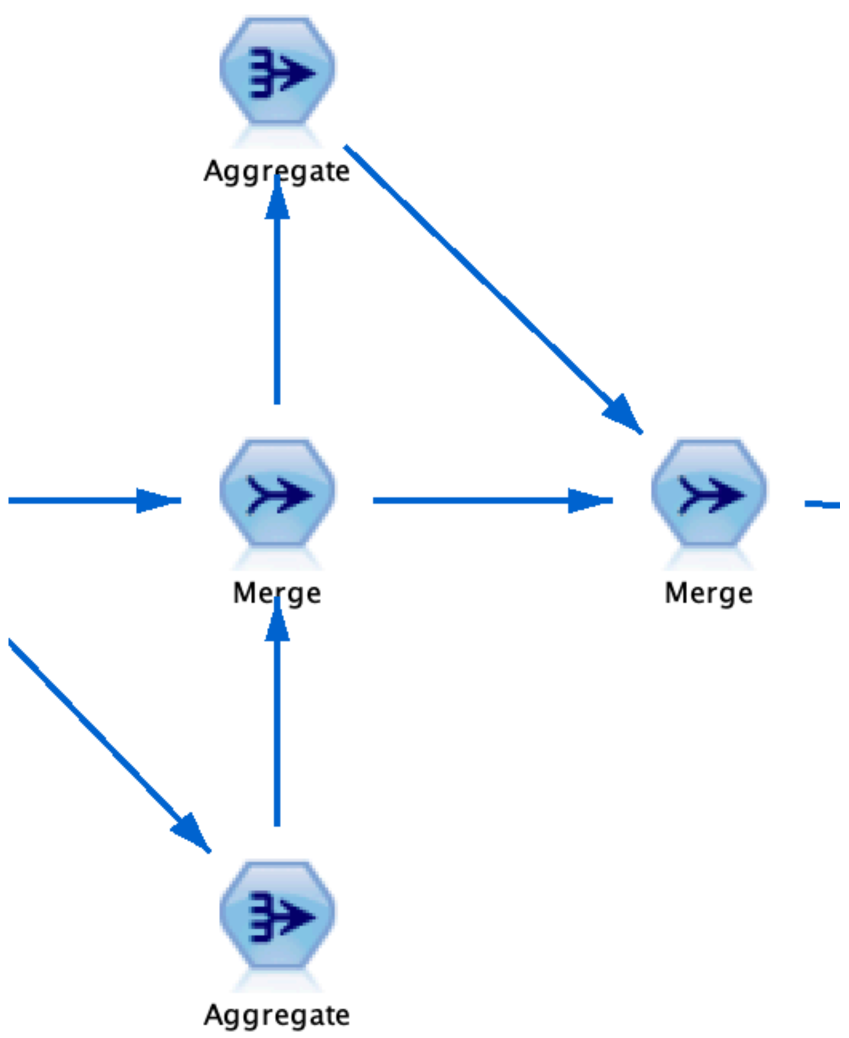
Understand who the most active users are and see what their occupation is: when there are many active users in the same workplace, they could be given discounts if they invite their colleagues to also join the rating platform

Business Goal 3: Determine where the most active users work

- First, we must understand who the active users are. We count every occurrence of the user id in the table through the aggregate node.

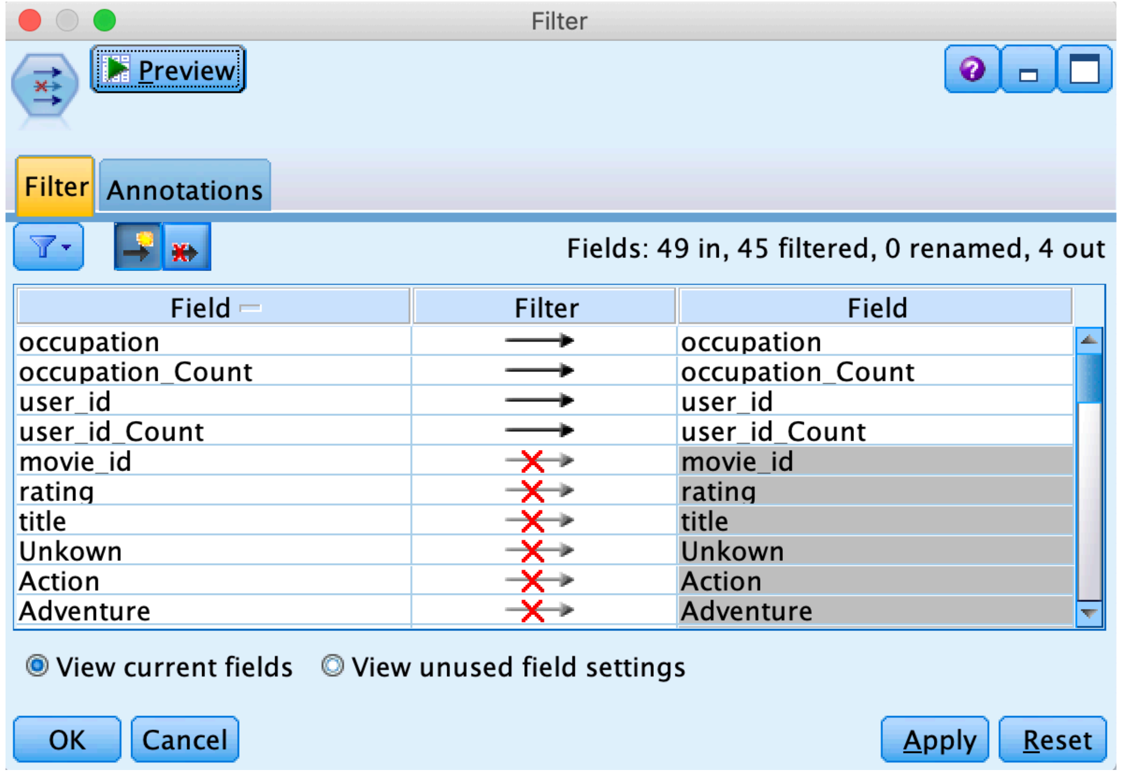
- Moreover, we must also understand how many people work in a certain workplace. Here, we also use an aggregate node to count every occurrence of an occupation in the table.

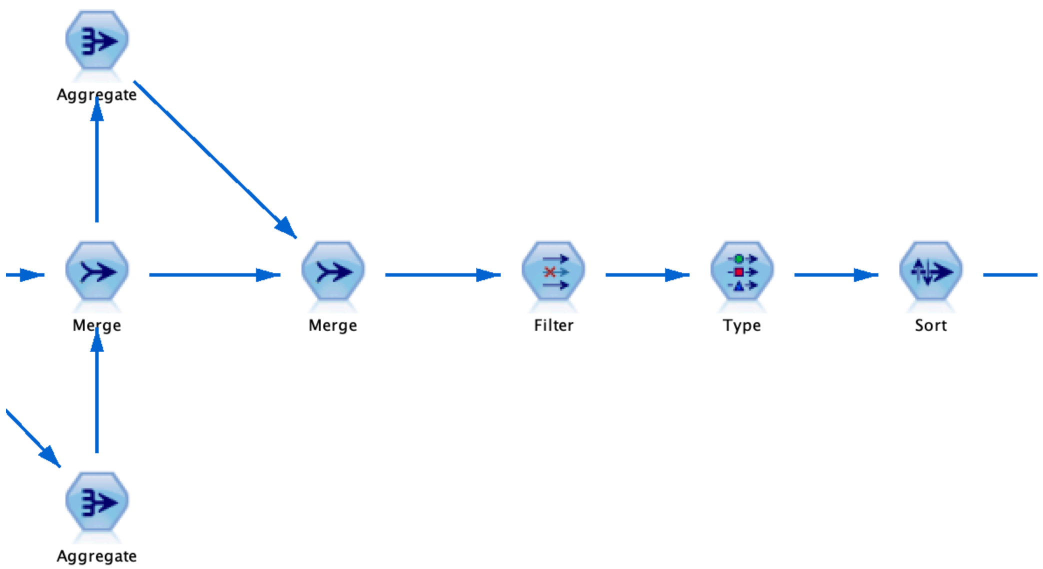
- The files are then merged together: active users on user-id and workplace on occupation



Business Goal 3:
Determine where the most active users work

- Next, everything but the following four columns are removed from the dataset through the filter node:
    - Occupation
    - Occupation_count
    - User_id
    - User_id_count
- Then, occupation type is set to nominal through the type node.
- The table is then sorted on user_id_count (descending)





Business Goal 3: Determine where the most active users work

- Every user has multiple entries (since there used to be multiple rows with different ratings). Since these columns have been deleted, the table contains many duplicate rows.

- Through the ‘distinct’ node, we ‘create a composite record for each group’ based on the user_id. Now, every user_id only has one row in the table and the most active user is ranked at the top:

	occupation	occupation_Count	user_id	user_id_Count
1	healthcare	2774	405	737
2	healthcare	2774	655	684
3	educator	8932	13	635
4	educator	8932	450	539
5	student	21852	276	517
6	student	21852	416	492
7	engineer	7900	537	489
8	student	21852	303	483
9	student	21852	393	447
10	executive	3310	181	434
11	program...	7574	279	433
12	student	21852	429	413
13	lawyer	1339	846	405
14	administr...	7392	7	403
15	student	21852	94	399
16	program...	7574	682	398
17	writer	5466	293	387
18	entertain...	2084	92	387
19	program...	7574	222	386

Business Goal 3: Determine where the most active users work

- A filter node is used to remove ‘record count’ from the dataset.

- We can now add a distribution graph and show where all the users work:

Value	Proportion	%	Count
administrator		8.45	77
artist		3.07	28
doctor		0.66	6
educator		9.88	90
engineer		7.03	64
entertainme...		1.98	18
executive		3.4	31
healthcare		1.65	15
homemaker		0.77	7
lawyer		1.32	12
librarian		5.49	50
marketing		2.85	26
none		0.99	9
other		11.42	104
programmer		7.03	64
retired		0.22	2
salesman		1.21	11
scientist		3.4	31
student		21.51	196
technician		2.85	26
writer		4.83	44

Business Goal 3: Determine where the most active users work

- However, we are interested in the most active users. Therefore, we will select the top 10% of the dataset. Since there are 911 records, we look at the 91st entry and see that the user count is 244. Therefore, we will use a select node to include all the entries that have a user_id_count of 244 or higher. The distribution now looks as follows:

Value	Proportion	%	Count
administrator		8.79	8
doctor		1.1	1
educator		8.79	8
engineer		8.79	8
entertainme...		3.3	3
executive		4.4	4
healthcare		2.2	2
lawyer		2.2	2
librarian		6.59	6
marketing		2.2	2
none		1.1	1
other		8.79	8
programmer		5.49	5
salesman		1.1	1
student		24.18	22
technician		2.2	2
writer		8.79	8

Maaike

Business Goal 3: Determine where the most active users work

- Even though we can see where the top 10% of active users work, it does not yet show where the most active users (1%) work. There is still a large difference between the 1st percentage and 10th percentage of active users. Therefore, we will show this distinction within a graph.

- Right now, the user only has an id or a count of how many times they have rated a movie. However, if we now bin the users, they will either be binned on user_id or their count. User_id does not take into account who the most active user is. The count will not be able to create bins of equal distance and equal users. Therefore, this step will need to be completed in two-fold.

- First, a derive node will be used. The derive node will count whenever the user_id is higher than 0 (which is always), ensuring that every entry receives their own rank. (This could have been done through binning with ranking order on percentile, but whenever a user_id_count was equal, one entry was ignored)

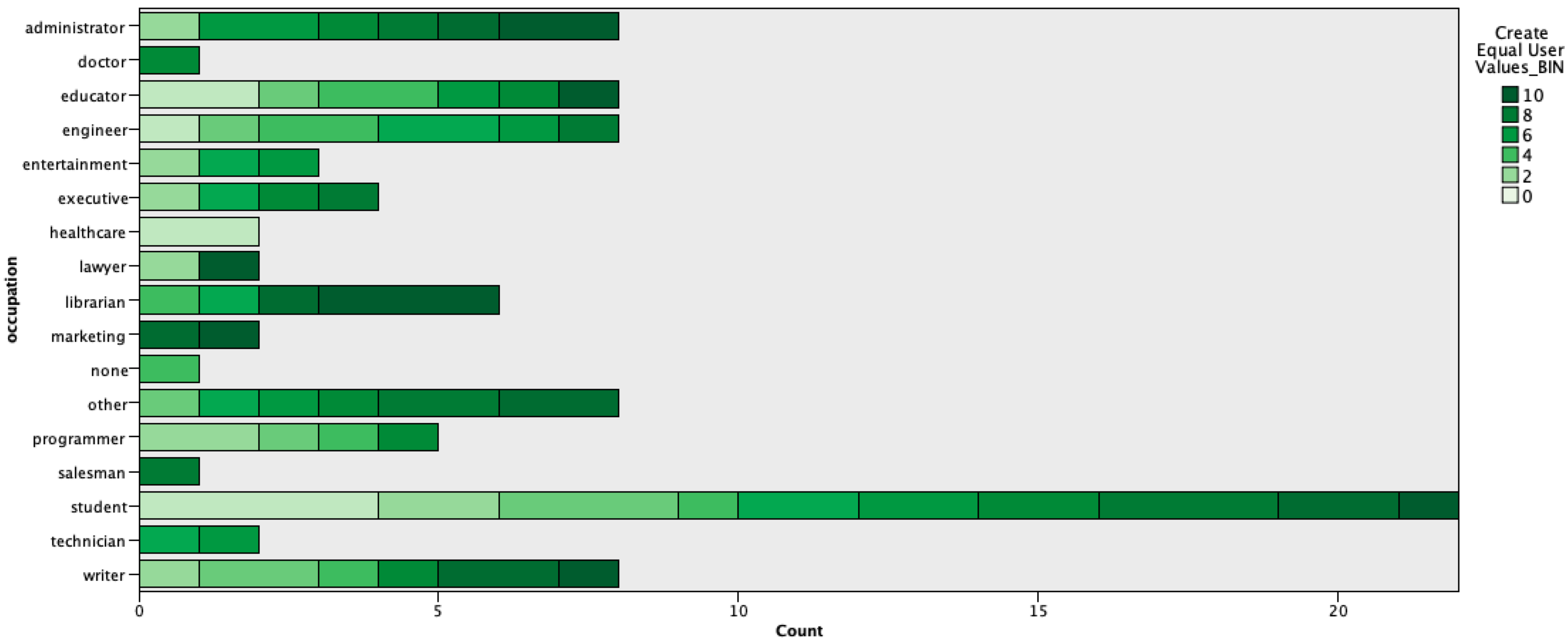
Business Goal 3: Determine where the most active users work

- We now have a ranking (variable is called Create Equal User Values), which allows us to create bins of equal amounts of users based on their rating activism.

- The bin node is used. It will bin the Create Equal User Values node into 10 bins, meaning every percentage will be shown (the top 10% of users is now again divided into 10 bins).

- Through the distribution graph node, the occupation will be shown with a color overlay of the binned users. The graph has a proportional scale in order to show the different bins better

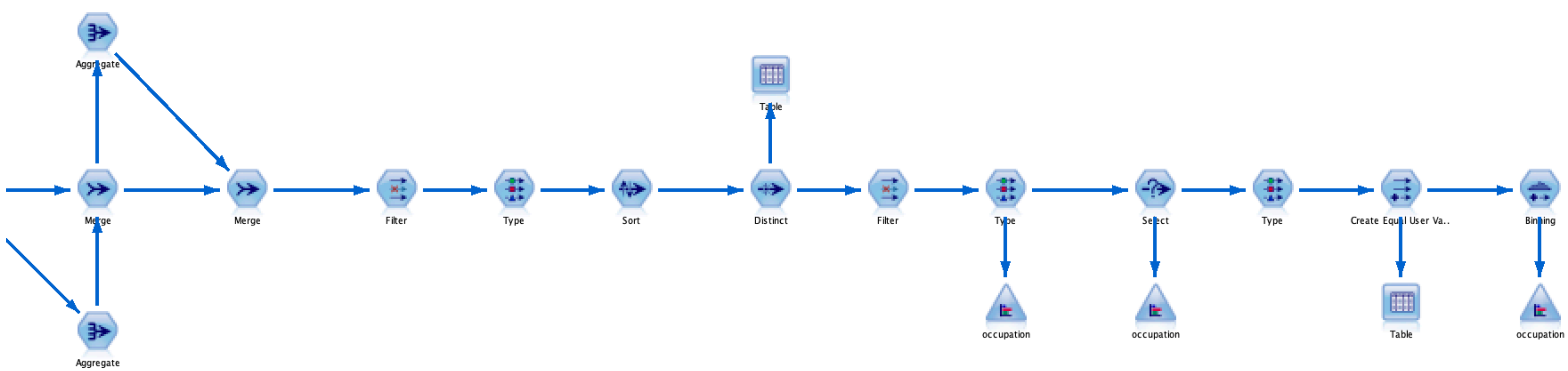
Business Goal 3: Determine where the most active users work



- Here, we can see the percentage of users that work in a specific workplace. The light green indicates that the user is very active. The darker the green gets, the less active the user is.

Maaike

Business Goal 3:
Determine where the most active users work



- The final setup for business goal 3 can be seen above. Of course, the data preparation will be added before this string in order to provide the complete result.

Business Goal 4: Recommender system

Recommend a movie genre that a user normally doesn’t watch but might like, by identifying users with similar tastes using the Pearson correlation.

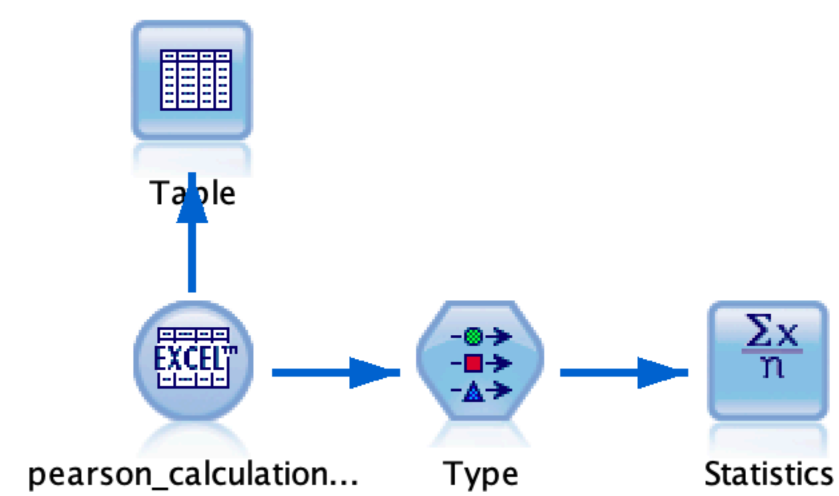
How would user_id #34 rate the Sci-fi movie genre (based on how similar users rated Sci-fi)?

Charles

Business Goal 4: Recommender system

- To recommend a new movie genre to someone, we must calculate the Pearson coefficient for a particular user. The pearson coefficient tells us a user's "nearest neighbors", that is, how similar other user's tastes are to the chosen person.
- First, we prepare the data to show just the normalized average rating per genre for each user
- Next, we use the Statistics node to calculate the pearson coefficient

user_id	2.000000	3.000000	4.000000	5.000000
Action_ave	0.178	-0.167	-0.448	-0.021
Adventure_ave	0.711	0.548	-0.823	0.078
Animation_ave	$null$	$null$	$null$	0.605
Childrens_ave	-0.955	$null$	$null$	-0.771
Comedy_ave	0.178	-0.369	0.677	-0.176
Crime_ave	0.156	0.048	0.427	0.725
Documentary_ave	$null$	2.048	0.677	$null$
Drama_ave	0.206	-0.043	0.177	-0.497
Fantasy_ave	-0.622	$null$	$null$	-0.664
Film-Noir_ave	0.878	-0.452	$null$	1.836
Horror_ave	-0.622	-0.552	-0.323	-0.628
Musical_ave	-0.622	-0.952	0.677	0.169
Mystery_ave	-0.122	0.229	-0.323	-0.164
Romance_ave	0.503	0.448	0.010	-0.848
Sci-Fi_ave	0.128	-0.202	-0.490	0.351
Thriller_ave	-0.039	-0.429	-0.414	-0.217
War_ave	0.045	-0.152	0.177	0.050
Western_ave	$null$	$null$	$null$	-0.664



$$\mathrm{PC}(u,v) = \frac{\sum\limits_{i \in \mathcal{I}_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum\limits_{i \in \mathcal{I}_{uv}} (r_{ui} - \bar{r}_u)^2 \sum\limits_{i \in \mathcal{I}_{uv}} (r_{vi} - \bar{r}_v)^2}}.$$

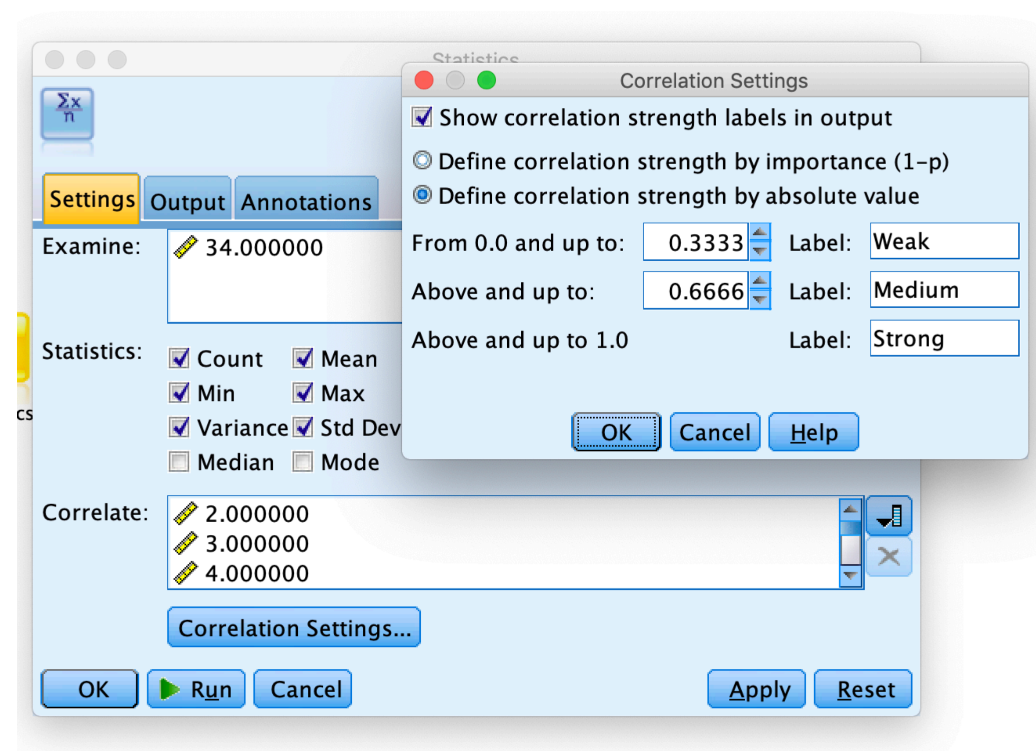
Charles

Business Goal 4: Recommender system

- We calculate the pearson coefficient for a particular user. We chose **user_id #34**, because this person has 6 null rating values out of 18. This is optimal because they have enough data to calculate an accurate pearson coefficient, but enough nulls for us to give them a prediction.

user_id	2.000000	3.000000	4.000000	5.000000
Action_ave	0.178	-0.167	-0.448	-0.021
Adventure_ave	0.711	0.548	-0.823	0.078
Animation_ave	$null$	$null$	$null$	0.605
Childrens_ave	-0.955	$null$	$null$	-0.771
Comedy_ave	0.178	-0.369	0.677	-0.176
Crime_ave	0.156	0.048	0.427	0.725
Documentary_ave	$null$	2.048	0.677	$null$
Drama_ave	0.206	-0.043	0.177	-0.497
Fantasy_ave	-0.622	$null$	$null$	-0.664
Film-Noir_ave	0.878	-0.452	$null$	1.836
Horror_ave	-0.622	-0.552	-0.323	-0.628
Musical_ave	-0.622	-0.952	0.677	0.169
Mystery_ave	-0.122	0.229	-0.323	-0.164
Romance_ave	0.503	0.448	0.010	-0.848
Sci-Fi_ave	0.128	-0.202	-0.490	0.351
Thriller_ave	-0.039	-0.429	-0.414	-0.217
War_ave	0.045	-0.152	0.177	0.050
Western_ave	$null$	$null$	$null$	-0.664

- We set up the statistics node like so to see which other users have similar movie genre tastes as user #34. We consider the following:
  - 0.0–0.333 = Weak relationship
  - 0.333–0.666 = Moderate relationship
  - 0.666–1.0 = Strong relationship



Charles

Business Goal 4: Recommender system

- Looking at the correlation results, we take the top 10 highest pearson coefficients to identify our 10 most similar neighbors to user #34

- Due to the limitations to the SPSS statistics node (can't sort or connect to another node), we export the data into Excel to sort the top 10 neighbors and find their rating for Sci-Fi_ave

- The 10 neighbors of user 34, from strongest to weakest, are: 162, 74, 438, 113, 674, 634, 772, 117, 76, 566

34.000000

Statistics

Count	12
Mean	-0.000
Min	-2.651
Max	1.349
Range	4.000
Variance	1.765
Standard Deviation	1.328
Standard Error of Mean	0.383

Pearson Correlations

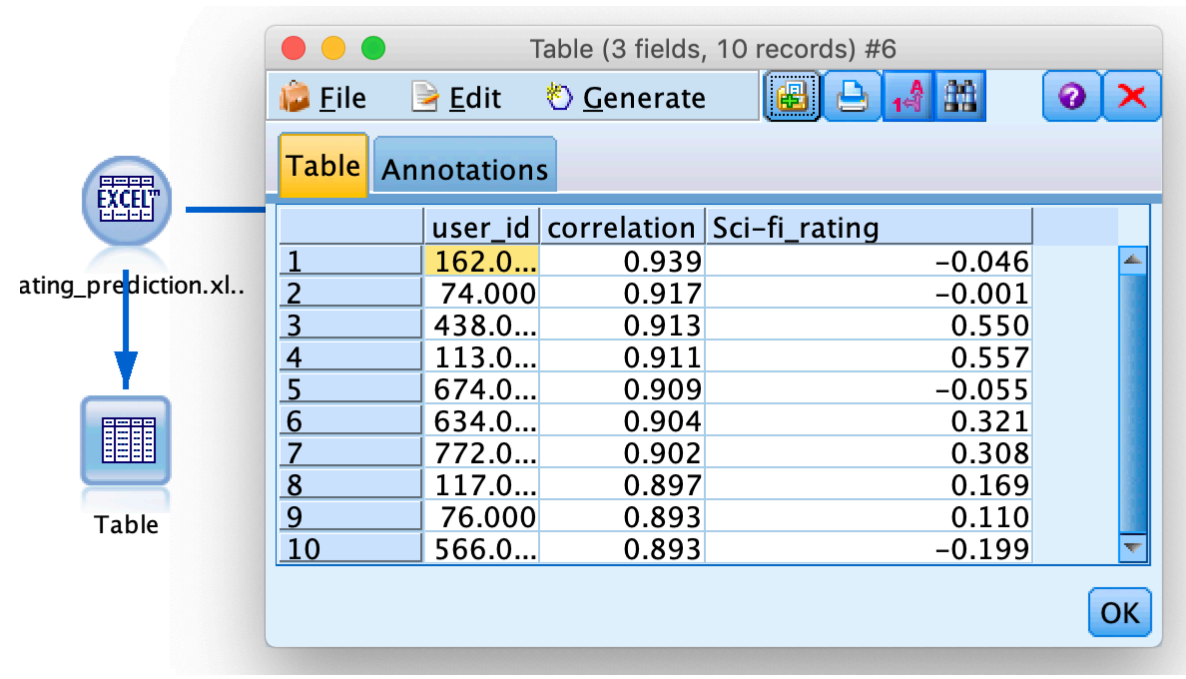
2.000000	0.843	Strong
3.000000	0.785	Strong
4.000000	-0.324	Weak
5.000000	0.320	Weak
6.000000	0.509	Medium
7.000000	0.073	Weak
8.000000	0.385	Medium
9.000000	-0.251	Weak
10.000000	0.381	Medium
11.000000	0.598	Medium
12.000000	0.344	Medium
13.000000	0.600	Medium
14.000000	0.657	Medium
15.000000	0.742	Strong
16.000000	-0.338	Medium
17.000000	-0.208	Weak
18.000000	0.354	Medium
19.000000	-0.443	Medium
20.000000	0.286	Weak
21.000000	0.246	Weak
22.000000	0.178	Weak
23.000000	0.475	Medium
24.000000	0.662	Medium
25.000000	0.514	Medium
26.000000	0.415	Medium
27.000000	0.285	Weak
28.000000	0.472	Medium
29.000000	0.696	Strong
30.000000	0.160	Weak
31.000000	0.160	Weak
32.000000	0.706	Strong
33.000000	0.506	Medium
35.000000	0.520	Medium
36.000000	0.184	Weak
37.000000	0.632	Medium
38.000000	-0.193	Weak

Charles

Business Goal 4: Recommender system

- We want to know, how would user 34 like the sci-fi genre? After all, they haven't rated it.

  To do so, we prepare and input the data into SPSS again, showing only the top 10 neighbors of user #34, their pearson correlation, and their average rating for the Sci-fi genre.



	user_id	correlation	Sci-fi_rating
1	162.0...	0.939	-0.046
2	74.000	0.917	-0.001
3	438.0...	0.913	0.550
4	113.0...	0.911	0.557
5	674.0...	0.909	-0.055
6	634.0...	0.904	0.321
7	772.0...	0.902	0.308
8	117.0...	0.897	0.169
9	76.000	0.893	0.110
10	566.0...	0.893	-0.199

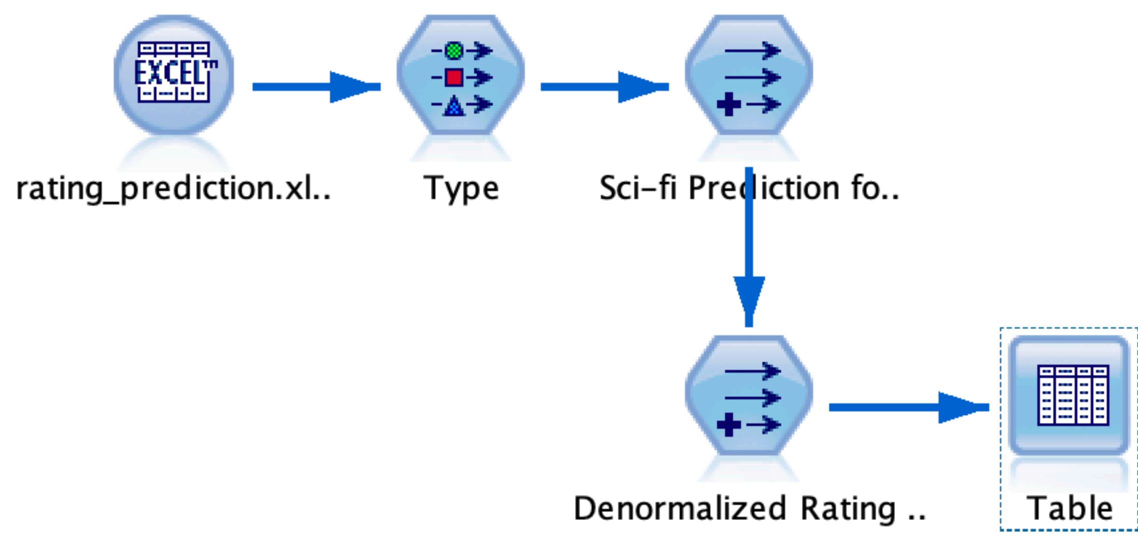
- We use the equation displayed here to predict the rating that user #34 would give to the Sci-fi genre, based on how his top 10 neighbors rated the Sci-fi genre.

$$\hat{r}_{ui} = \frac{\sum\limits_{v \in \mathcal{N}_i(u)} w_{uv} r_{vi}}{\sum\limits_{v \in \mathcal{N}_i(u)} |w_{uv}|}.$$

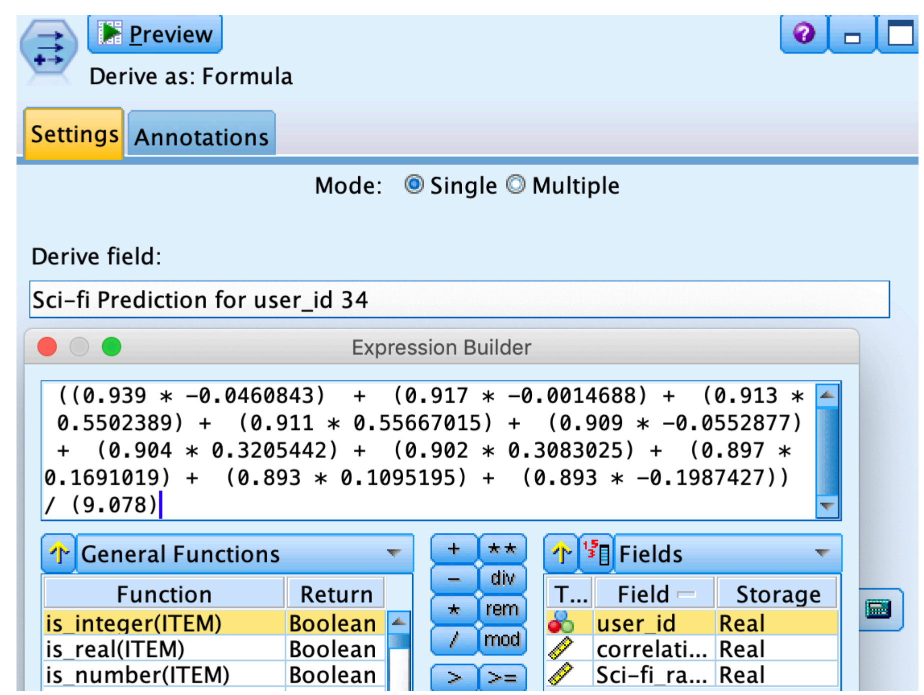
Charles

Business Goal 4: Recommender system

- We perform the equation using the Derive node



- The result, which is the the normalized prediction for how user 34 would rate the Sci-fi genre is, 0.171.



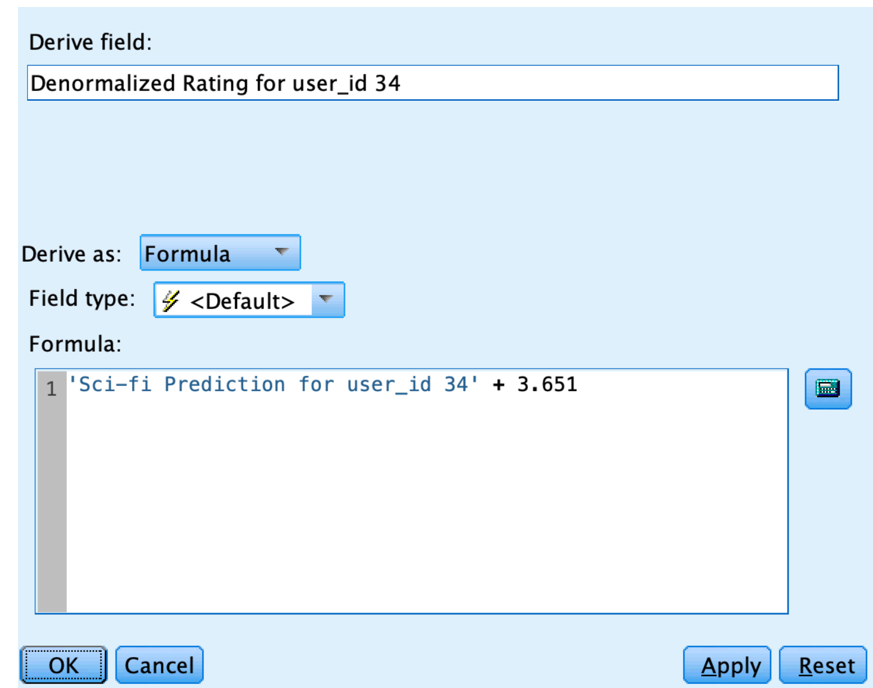
Charles

Business Goal 4: Recommender system

- We de-normalize the prediction by using another Derive node and adding user #34's average rating for all genres, 3.651, back to their normalized prediction.

- The final predicted rating for Sci-fi for user 34 is **3.822!**

- This can be rounded to 4. This is rating, 4 (out of max 5) can be considered **strong**. Thus, we can now confidently recommend Sci-fi movies to user #34, despite **this person having never rated this genre to begin with!!**



user_id	correlation	Sci-fi_rating	Sci-fi Predict...	Denormalized Rating for user_id 34
162	0.939	-0.046	0.171	3.822
74	0.917	-0.001	0.171	3.822
438	0.913	0.550	0.171	3.822
113	0.911	0.557	0.171	3.822
674	0.909	-0.055	0.171	3.822
634	0.904	0.321	0.171	3.822
772	0.902	0.308	0.171	3.822
117	0.897	0.169	0.171	3.822
76	0.893	0.110	0.171	3.822
566	0.893	-0.199	0.171	3.822

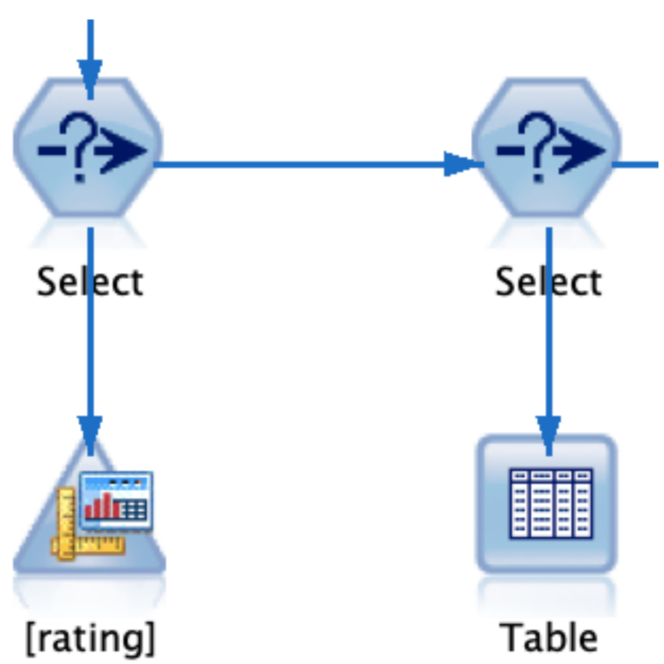
Business Goal 5: Determine the most popular movie for a selected age group

Find out the most popular movie for certain age group and create recommendations for the Theme park for making new attractive rides. Help in attracting a different age group to these theme parks using the dataset.

Business Goal 5: Determine the most popular movie for a selected age group

- From the analysis done for the Business Goal 1, we found out the popular age group (19-40). We decided to go forward with the same age group because we wanted to find interesting movies for making this age group also interested in the theme park rides.

- We found out the movies that are interesting for the chosen age group.



user_id	movie_id	rating	title
3	355	0.048	Sphere (1998)
3	345	0.048	Deconstructing Harry (1997)
3	340	2.048	Boogie Nights (1997)
3	260	1.048	Event Horizon (1997)
3	268	0.048	Chasing Amy (1997)
3	354	0.048	Wedding Singer, The (1998)
3	351	0.048	Prophecy II, The (1998)
3	307	0.048	Devils Advocate, The (1997)
3	331	1.048	Edge, The (1997)
3	299	0.048	Hoodlum (1997)
3	329	1.048	Desperate Measures (1998)
3	320	2.048	Paradise Lost: The Child M...
3	346	2.048	Jackie Brown (1997)
3	318	1.048	Schindlers List (1993)
3	322	0.048	Murder at 1600 (1997)
3	344	1.048	Apostle, The (1997)
3	321	2.048	Mother (1996)
3	334	0.048	U Turn (1997)
3	327	1.048	Cop Land (1997)
3	350	0.048	Fallen (1998)
3	328	2.048	Conspiracy Theory (1997)
3	343	0.048	Alien: Resurrection (1997)
3	342	1.048	Man Who Knew Too Little, ...
3	348	1.048	Desperate Measures (1998)
3	181	1.048	Return of the Jedi (1983)
3	349	0.048	Hard Rain (1998)
3	339	0.048	Mad City (1997)
3	271	0.048	Starship Troopers (1997)
3	303	0.048	Ulees Gold (1997)
3	347	2.048	Wag the Dog (1997)
4	324	0.677	Lost Highway (1997)
4	362	0.677	Blues Brothers 2000 (1998)
4	258	0.677	Contact (1997)
4	50	0.677	Star Wars (1977)
4	303	0.677	Ulees Gold (1997)
4	354	0.677	Wedding Singer, The (1998)
4	300	0.677	Air Force One (1997)
4	360	0.677	Wonderland (1997)
4	294	0.677	Liar Liar (1997)
4	329	0.677	Desperate Measures (1998)
4	327	0.677	Cop Land (1997)
4	359	0.677	Assignment, The (1997)
4	361	0.677	Incognito (1997)
4	301	0.677	In & Out (1997)
9	487	0.418	Roman Holiday (1953)
9	691	0.418	Dark City (1998)

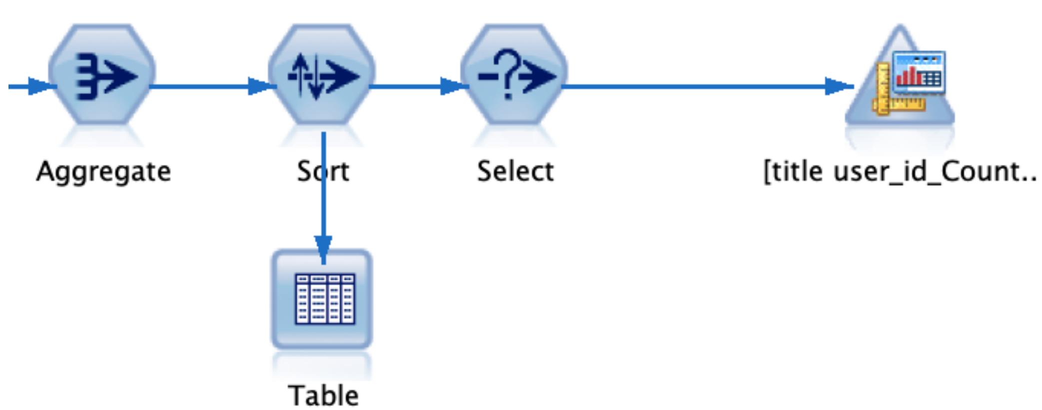
Pammu

Business Goal 5: Determine the most popular movie for a selected age group

- Then we found out the number of users who gave high ratings for each movies, to find the most popular movies from the chosen age group.

- After sorting and selecting the top ones, we got a very interesting graph showing some interesting results.

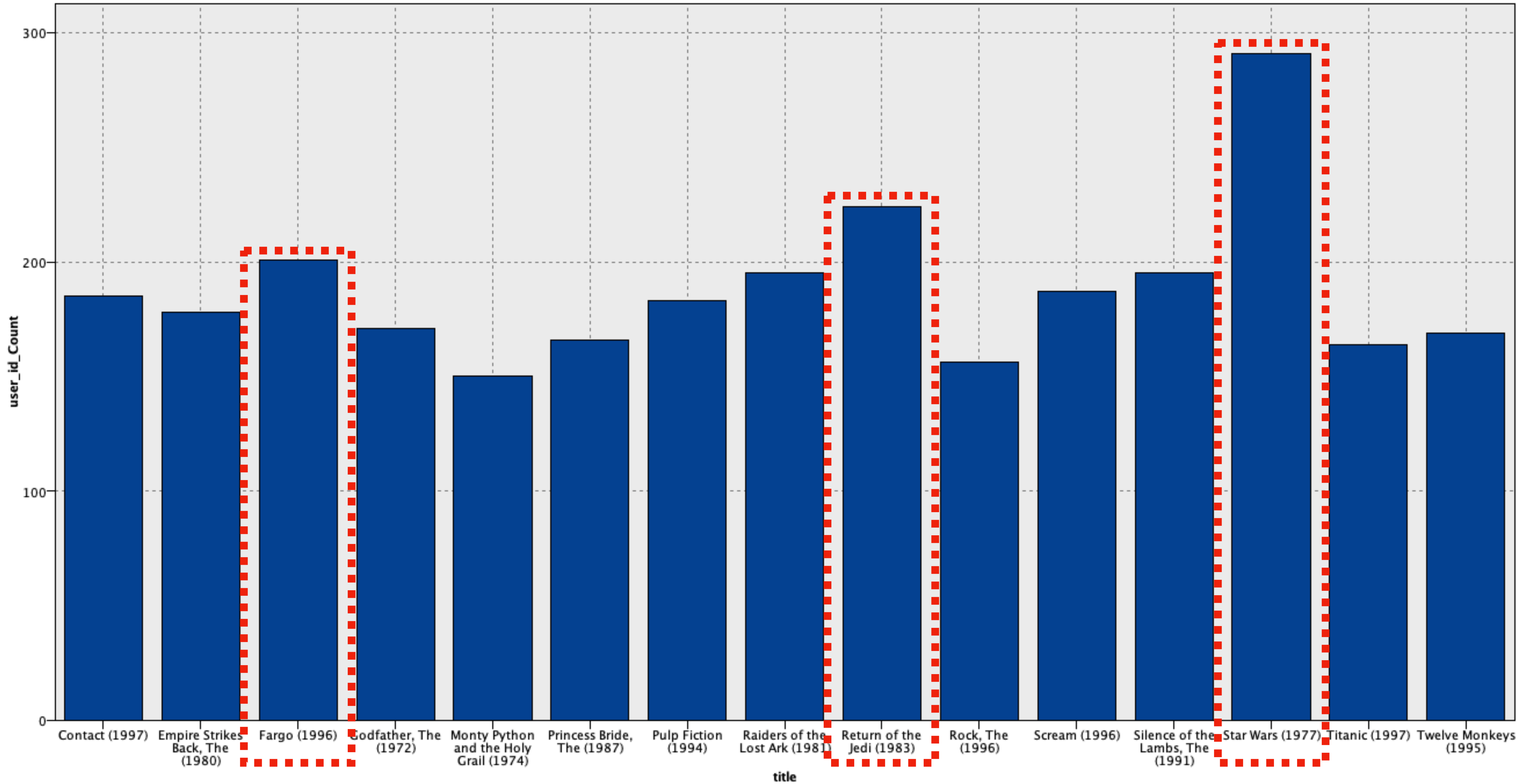
- People from the age group of 19-40 still love movies from the 18th century and this would be a valuable information for building rides for this age group.



user_id_Count	movie_id	title
291	50	Star Wars (1977)
224	181	Return of the Jedi (1983)
201	100	Fargo (1996)
195	98	Silence of the Lambs, The (1991)
195	174	Raiders of the Lost Ark (1981)
187	288	Scream (1996)
185	258	Contact (1997)
183	56	Pulp Fiction (1994)
178	172	Empire Strikes Back, The (1980)
171	127	Godfather, The (1972)
169	7	Twelve Monkeys (1995)
166	173	Princess Bride, The (1987)
164	313	Titanic (1997)
156	117	Rock, The (1996)
150	168	Monty Python and the Holy Grai...
148	64	Shawshank Redemption, The (1...
143	12	Usual Suspects, The (1995)
143	79	Fugitive, The (1993)
141	318	Schindlers List (1993)
138	22	Braveheart (1995)
137	222	Star Trek: First Contact (1996)
137	96	Terminator 2: Judgment Day (1...
134	300	Air Force One (1997)
132	237	Jerry Maguire (1996)
132	210	Indiana Jones and the Last Crus...
129	286	English Patient, The (1996)
129	475	Trainspotting (1996)
128	176	Aliens (1986)
127	183	Alien (1979)
126	294	Liar Liar (1997)
126	195	Terminator, The (1984)
124	11	Seven (Se7en) (1995)
124	89	Blade Runner (1982)
123	204	Back to the Future (1985)
123	151	Willy Wonka and the Chocolate ...
122	69	Forrest Gump (1994)
120	302	L.A. Confidential (1997)
118	216	When Harry Met Sally... (1989)
117	268	Chasing Amy (1997)
116	121	Independence Day (ID4) (1996)
111	196	Dead Poets Society (1989)
109	257	Men in Black (1997)
108	191	Amadeus (1984)
107	202	Groundhog Day (1993)
106	357	One Flew Over the Cuckoos Nes...
104	483	Casablanca (1942)
102	276	Leaving Las Vegas (1995)

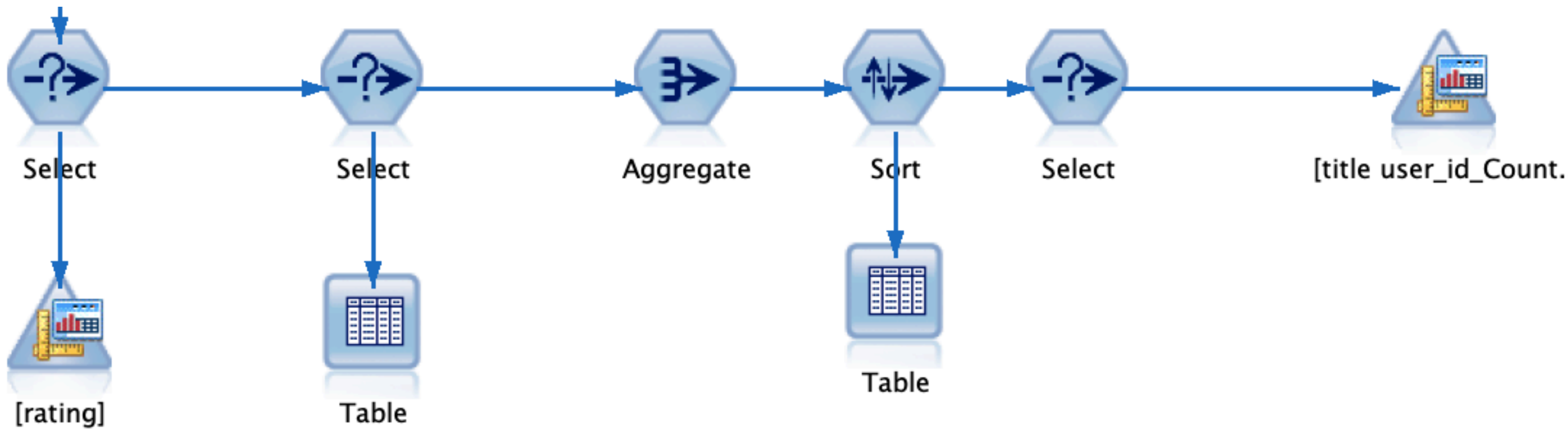
Pammu

Business Goal 5:
Determine the most popular movie for a selected age group



Pammu

Business Goal 5:
Determine the most popular movie for a selected age group



- The final setup for business goal 5 can be seen above. Of course, the data preparation will be added before this string in order to provide the complete result.

Pammu