Survey Methodology

MOVIES GENRES

Introduction

This report will go through the steps and actions we did at each level. The five steps include: the questionnaire, samples and pretests, data collection, expectation vs. reality, and data analysis.

The questionnaire phase

Introduction

Genres matter in both cinema and television because they develop expectations. When people go to the theatre or sit down to watch television, the audience chooses a specific experience. Consider how, when asked whether you want to see a movie, one of the first questions you will ask is, "What kind of movie?" Your query about the genre of film. And here is where the poll to determine audience film preferences kicks in.

Main survey questions

Surveys are made to elicit knowledge and opinions from a large population. Researchers use carefully constructed survey questions to elicit particular types of replies that can aid in understanding attitudes, opinions, behaviours, and other crucial data relevant to their study objectives. People can readily comprehend and respond to the questions. The questions in the home segment are different from the questions in the cinema section depending on whether the test-taker sees films at home or in a cinema. There are inquiries to see whether the person is responding haphazardly or not. The individual may choose not to answer some questions.

Response options

There are a variety of questions, including multiple-choice, rating scales, tick boxes, and open-ended inquiries.

There are also some personal questions left at the end so that the participants can get comfortable first with the survey.

There are 43 questions:

Cinema: 22

- MCQ: 15
- Rating: 2
- Checkbox: 2 (1: Optional)
- Open: 3

Home: 21

- MCQ: 10
- Rating: 2
- Checkbox: 3 (1: Optional)
- Open: 6 (1: Optional)

Sampling and pretest phase

Introduction

It is vital to identify all the components needed for concluding the sampling process in the third phase of this research investigation. To guarantee the accuracy and authenticity of the data gathered, this phase entails rigorous preparation and execution of the sampling technique. To get accurate and significant data that may be extrapolated to the target population, a well-executed sampling procedure is crucial.

2.1 – Defining the target population

So there are certain **characteristics** we are going to define for the target population:-

- **Age**: We believe that movies appeal to people of all ages, so our poll attempts to collect opinions on movie preferences from a wide spectrum of people. To achieve a complete representation of ideas and opinions, we have decided not to impose any specific age limitations on the group being polled.

- **Gender**: We want responses from people of all gender identities for our survey. This open-minded strategy enables us to examine the preferences and viewpoints of both men and women, which may offer insightful information about the kinds of films that appeal to each gender.
- **Location**: As it is crucial to gather a wide range of opinions and experiences relating to films, our survey strives to include respondents from a variety of geographic places. Even though we are concentrating on Alexandria, Egypt's college students, we understand the value of hearing from people in other places. As a result, we intend to use social media to increase our audience and attract participants from around the world.

2.2 – Choosing the Sampling Method

There are several things to think about when it comes to sample techniques. However, we have decided that the best strategy for our investigation is **random sampling**. This choice was made since we don't want any age group to predominate the sample and there is an equal distribution of first, second, third, and fourth-grade pupils in the population. We can lessen the effects of potential biases by utilizing random sampling, which improves the generalizability of our findings to a larger population. To achieve this, we will select participants at random from the selected group and distribute the questionnaire to them using Google Forms.

2.3 – Determining the sample size

Although there are other ways to determine sample size, for this study we are limited to a particular number of participants—roughly 400, which equates to the total number of students in the other groups. We think that a sample size of 400 is sufficient for our investigation, despite the fact that such a restriction may have disadvantages in some situations.

Validating the entire study process is one of the third phase's main goals. This involves evaluating the analytical approaches, data collection strategies, sampling protocols, and research tools to make sure they are relevant and useful for the study issue. In this aspect, the pretest phase is particularly crucial because it aids in identifying any potential issues or flaws in the research design, allowing researchers to make the required adjustments prior to starting the real data collection.

3.1 - Selecting the group of participants for the test

Our research has shown that the best sample size for evaluating a survey on a particular group is between 5% and 10% of the main population. We set a sample size of 30 people to guarantee that we have a representative sample because our primary group is made up of 300 to 400 people. In a manner similar to how we chose the main group, we randomly selected this sample. We can gain insights and make wise decisions by getting input from this sample.

3.2 - Administering the test of the survey

We will distribute the questionnaire to our sample using "Google Forms," a free Internet tool. With this strategy, we may shorten the data-gathering process and efficiently gather responses. After each session, we will include a question to get feedback from participants. For example, we might ask if any concepts were misunderstood or if any questions might have been phrased more effectively. Of course, this question will be open-ended to ensure that we receive objective input from the participants.

3.3 - Analyzing the results and determining modifications

We will now examine the survey response data to identify any changes that need to be made to the survey or study's design. This stage is crucial for guaranteeing the validity and dependability of the study's conclusions. After the responses have been gathered, the data will be analyzed to see whether any changes to the questionnaire are required. These changes will be made in light of the input that has been received.

Modifications could involve altering the phrasing or format of the questions, including or excluding certain ones, or changing the demographics or survey sample size. This stage is essential to collecting high-quality data and assuring the validity and dependability of the study.

Collecting the Data Phase

Introduction

The objective of this phase is to obtain a clean dataset that is free from careless responses, missing values and inconsistencies while being well-framed and easy to read. As part of the data collection process, it is essential to ensure that the data obtained is clean and free from any errors or inconsistencies. In this phase, we will be focusing on the pre-analysis of the collected data to obtain a clean dataset that is well-framed, easy to read and understandable. This will involve reviewing the survey responses to identify any careless answers, missing values or inconsistencies. Another important aspect of data cleaning is identifying and dealing with missing values. Missing values can occur when a participant fails to answer a question or if there was an error in the data collection process. It is crucial to handle missing values appropriately to ensure that the analysis is accurate and unbiased. Finally, the data set needs to be well-framed, which means that it is presented in a way that is easy to read and understand. Well-framed data helps to ensure that the insights gained from the analysis are clear and relevant to the research questions at hand.

How the data was framed

To ensure that our data is clearly framed and simple to understand, we utilized the Python programming language and the Pandas library. In order to determine the frequency with which each checkbox was selected, we divided several columns that represented checkbox questions in our form. We also changed the names of the columns on the form that were named after the questions. We made an effort to keep it brief and to describe the data.

TECHNIQUES USED

1. Renaming Columns
2. Dividing Checkbox (Best day, Best time)
3. Dividing Checkbox (Favorite Genre)
4. Pilot testing: Identifying issues before Full Implementation, we have already performed it on the first 70 responses to check if we refine question wording or instructions, to ensure that the survey is understandable and relevant to the target population and improve the quality and validity of the survey instrument.
5. Validity and Reliability: To check reliability, we have used **Cronbach's alpha** (A measure of internal consistency reliability that assesses the degree to which a set of survey questions measure the same construct. questions are measuring the same construct, It ranges from 0 to 1, with higher values indicating greater internal consistency). To achieve the highest internal consistency reliability (Cronbach's alpha), we have used the corr() function to get the

highest correlated features. **Conclusion**: The highest value of alpha we have achieved is: **0.803** which is acceptable.

6. Test-Retest: To check the validity of our responses, we applied this technique on sample size = 30 and we compared their responses at different times for the same person (same id). we consider our question to check validity. we **conclude** that only 2 responses of our 30-sample have answered different answers between each one he had answered it, so we have achieved validity.

   In the case of Test-Retest we have gone through three steps:

   First: We identified our variables, We used four features to calculate the reliability of our data

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Student_ID	Numeric	14	0	Student ID	None	None	15	Center	Nominal	Input
2	Student_Age	Numeric	2	0	Student Age	None	None	13	Center	Scale	Input
3	R11	Numeric	8	0	Action	{1, Least Fav...	None	9	Center	Ordinal	Input
4	R12	Numeric	8	0	Adventure	{1, Least Fav...	None	9	Center	Ordinal	Input
5	R13	Numeric	8	0	Comedy	{1, Least Fav...	None	9	Center	Ordinal	Input
6	R14	Numeric	8	0	Drama	{1, Least Fav...	None	9	Center	Ordinal	Input
7	R15	Numeric	8	0	Romance	{1, Least Fav...	None	9	Center	Ordinal	Input
8	R16	Numeric	8	0	Horror	{1, Least Fav...	None	9	Center	Ordinal	Input
9	R17	Numeric	8	0	Sci-fi or Fantancy	{1, Least Fav...	None	9	Center	Ordinal	Input
10	R18	Numeric	8	0	Musical	{1, Least Fav...	None	9	Center	Ordinal	Input
11	R19	Numeric	8	0	Family	{1, Least Fav...	None	9	Center	Ordinal	Input
12	R1	Numeric	8	0	Ranking Movies Genres	None	None	8	Center	Ordinal	Input
13	R21	Numeric	8	0	Cast	{1, Least Fav...	None	6	Center	Ordinal	Input
14	R22	Numeric	8	0	Genre	{1, Least Fav...	None	6	Center	Ordinal	Input
15	R23	Numeric	8	0	Director	{1, Least Fav...	None	6	Center	Ordinal	Input
16	R24	Numeric	8	0	Studio	{1, Least Fav...	None	6	Center	Ordinal	Input
17	R25	Numeric	8	0	Reviews	{1, Least Fav...	None	6	Center	Ordinal	Input
18	R26	Numeric	8	0	Quality of Story	{1, Least Fav...	None	6	Center	Ordinal	Input
19	R2	Numeric	8	0	Ranking the Factors Deciding Movies	None	None	8	Center	Ordinal	Input
20	Reasearching_Movies_Before_Watching	Numeric	8	0	Do you research movies before watching them?	{0, No}...	None	20	Center	Ordinal	Input
21	Where_You_Watch_Movies	Numeric	8	0	Where do you usually watch movies?	{1, Home}...	None	14	Right	Ordinal	Input
22	TOTAL	Numeric	40	0	Total	None	None	8	Center	Ordinal	Input

   Second: After identifying the variables we went to the data view interface to assign the individuals responses to make the analysis for our chosen features to calculate the reliability

	Student_ID	Student_Age	R11	R12	R13	R14	R15	R16	R17	R18	R19	R1	R21	R22	R23	R24	R25	R26	R2	Reasearching_Movies_Before_Watching
1	20221451429	20	2	2	4	5	5	3	5	2	4	32	5	4	2	3	4	5	23	2
2	20221451429	20	2	2	5	5	5	3	5	2	3	32	4	4	3	4	5	5	25	2
3	20221509866	20	5	5	5	5	3	3	4	2	3	35	5	4	3	4	5	5	26	2
4	20221509866	20	5	4	5	4	3	3	3	2	3	32	5	4	3	4	5	4	25	2
5	20221445288	20	4	5	3	3	4	5	3	1	3	31	5	3	1	1	5	4	19	2
6	20221445288	20	4	5	3	3	4	5	3	2	3	32	4	4	2	2	4	5	21	2
7	20221451449	20	1	4	2	1	2	1	3	5	4	23	4	2	1	1	5	5	18	2
8	20221451449	20	4	4	2	2	2	1	1	5	4	25	4	4	1	1	5	5	20	2
9	20221462492	20	5	4	4	5	3	5	5	2	2	35	5	5	4	4	4	5	27	1
10	20221462492	20	5	4	4	5	3	5	5	2	2	35	5	4	3	3	4	5	24	1
11	20221320560	20	1	2	4	3	4	1	3	3	4	25	4	3	1	3	4	4	19	1
12	20221320560	20	1	3	5	4	5	1	3	3	2	27	5	4	3	3	4	5	24	1
13	20221379966	20	5	5	3	4	4	3	5	3	4	36	5	5	4	3	4	5	26	2
14	20221379966	20	5	5	3	4	4	3	5	3	4	36	5	5	4	3	4	5	26	2
15	20221461977	19	2	3	2	2	3	4	1	1	1	19	3	3	1	1	2	3	13	2
16	20221461977	19	2	4	3	3	4	5	1	1	1	24	3	4	1	1	3	5	17	2
17	20221469438	20	5	4	5	2	1	5	3	3	2	0	4	5	2	1	4	5	21	1
18	20221469438	20	5	4	5	2	1	5	4	3	2	31	4	5	2	1	4	5	21	1
19	20221452375	20	3	5	3	3	3	4	5	3	5	34	5	5	5	3	3	5	26	1
20	20221452375	20	3	5	3	3	3	4	5	4	5	34	5	5	5	3	3	5	26	1
21	20221385544	20	3	4	2	3	2	3	5	2	2	26	4	5	3	2	5	5	24	1
22	20221385544	20	3	5	2	3	2	3	5	2	2	27	4	5	3	2	3	5	22	1
23	6	20	2	2	2	3	4	1	2	4	3	23	4	3	4	3	3	5	22	0
24	6	20	4	4	3	3	4	2	4	4	3	31	1	4	1	1	1	5	13	0
25	20221060922	19	3	4	1	5	1	4	1	1	3	23	4	5	3	3	4	5	24	1
26	20221060922	19	4	4	1	5	1	3	4	1	4	27	4	5	3	2	4	5	23	1
27	20221452894	20	5	3	5	4	4	1	2	2	4	30	5	5	1	3	3	5	22	2
28	20221452894	20	5	3	5	3	3	1	3	2	5	30	5	5	2	2	4	5	23	2

The columns are identifying our chosen questions:

1. Ranking the movie genres
2. Ranking factors deciding the movie
3. Do you research movies before watching?
4. Where do you usually watch movies?

Third: We calculate the reliability of our testing method which is Test-Retest Reliability

Reliability

Scale: ALL VARIABLES

Case Processing Summary

		N	%
Cases	Valid	42	100.0
	Excluded[a]	0	.0
	Total	42	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	N of Items
.803	20

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Action	158.14	381.882	.434	.793
Adventure	157.55	397.132	.314	.799
Comedy	157.95	389.071	.378	.796
Drama	157.81	387.865	.455	.795
Romance	157.98	401.536	.125	.804
Horror	158.64	398.333	.144	.804
Sci-fi or Fantancy	157.67	377.789	.589	.789
Musical	159.07	410.556	-.048	.809
Family	158.57	398.056	.214	.802
Ranking Movies Genres	131.79	270.075	.849	.749
Cast	157.21	388.709	.564	.794
Genre	157.19	399.573	.263	.801
Director	158.88	377.668	.636	.788
Studio	158.95	383.461	.493	.792
Reviews	157.43	398.934	.268	.801
Quality of Story	156.60	403.954	.330	.802
Ranking the Factors Deciding Movies	139.00	309.220	.765	.761
Do you research movies before watching them?	160.10	408.966	.016	.805
Where do you usually watch movies?	160.29	407.965	.102	.804
Total	106.79	180.514	.998	.758

	Student_ID	Student_Age	R11	R12	R13														
1	20221451429	20	2	2	4														
2	20221451429	20	2	2	5														
3	20221509866	20	5	5	5														
4	20221509866	20	5	4	5														
5	20221445288	20	4	5	3														
6	20221445288	20	4	5	3														
7	20221451449	20	1	4	2														
8	20221451449	20	4	4	2														
9	20221462492	20	5	4	4														
10	20221462492	20	5	4	4														
11	20221320560	20	1	2	4														
12	20221320560	20	1	3	5														
13	20221379966	20	5	5	3														
14	20221379966	20	5	5	3														
15	20221461977	19	2	3	2														
16	20221461977	19	2	4	3														
17	20221469438	20	5	4	5														
18	20221469438	20	5	4	5														
19	20221452375	20	3	5	3														
20	20221452375	20	3	5	3														
21	20221385544	20	3	4	2														
22	20221385544	20	3	5	2														
23	6	20	2	2	2														
24	6	20	4	4	3														
25	20221060922	19	3	4	1	5	1	4	1	1	3	23	4	5	3	3	4	5	24	1
26	20221060922	19	4	4	1	5	1	3	4	1	4	27	4	5	3	2	4	5	23	1
27	20221452894	20	5	3	5	4	4	1	2	2	4	30	5	5	1	3	3	5	22	2
28	20221452894	20	5	3	5	3	3	1	3	2	5	30	5	5	2	2	4	5	23	2

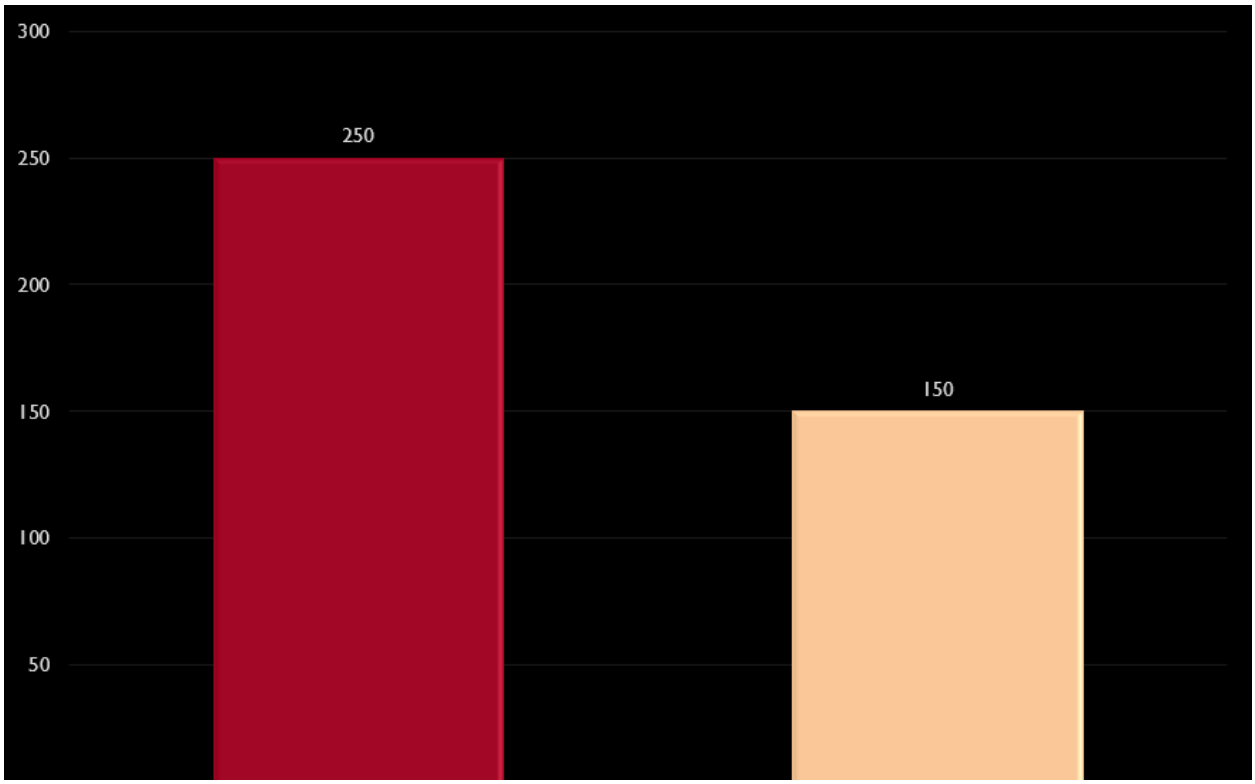
Expectation vs Real Phase

Introduction

Phase Five is a critical stage that focuses on evaluating the survey responses based on various metrics. This phase involves assessing the total number of answered surveys, identifying the number of surveys with carelessness-answered responses, determining the total number of non-answered surveys, and selecting the number of surveys that will be considered for analysis. **Our goal** is to provide explanations for the number of surveys chosen for analysis based on specific criteria. All of this highlights the importance of data quality, sample representativeness, statistical power requirements, and research objectives in determining the surveys that will be taken into consideration, ensuring reliable and meaningful results. Based on our proposal, we anticipate a response from a total of 400 participants in our survey. With this in mind, we are currently progressing towards presenting our findings

1 - Total number of answered surveys

Out of the total participants surveyed, our comprehensive analysis of the Google form questionnaire indicates that 250 participants responded, indicating a notable level of engagement and involvement from the target audience.



The response rate observed from this representation was 62.5%, indicating a moderate/high level of engagement from the participants, and 37.5% indicated the total percentage of unanswered surveys, which is important for determining the dataset's representativeness and the overall response rate. It gives information about participants' openness to sharing their ideas and opinions, demonstrating their level of interest and involvement in the survey.

2- Non-answered surveys

Non-answered surveys refer to those that were not completed by participants or had one or more questions left unanswered. This includes surveys with incomplete responses, skipped questions, or instances where respondents did not provide any input. Non-answered surveys contribute

to non-response bias and can impact the representativeness and generalizability of the findings. Out of the total participants surveyed, our comprehensive analysis of the Google form questionnaire indicates that 150 participants do not respond to our survey.

In survey analysis, keeping track of unanswered surveys is crucial because it enables the detection of potential biases and assesses the accuracy of the data. We believe that this is the cause of unanswered surveys. It is that none of the survey respondents had a financial incentive to participate. We feel that the percentage of unanswered surveys may have dropped if there had been financial support and incentives for the participants.

3 - Numbers of carelessness-answered surveys

Survey responses that show symptoms of inattention or a lack of effort from the participants are referred to as carelessly answered surveys. These responses could be answering all questions with the same response option, giving answers that are illogical or contradictory, or purposefully failing attention check questions. Carelessly completed surveys can jeopardize the accuracy of the information and the veracity of the analysis.

METHODOLOGY

To identify carelessness-answered surveys, various techniques were employed. These techniques included the implementation of attention check questions strategically placed within the survey to assess the

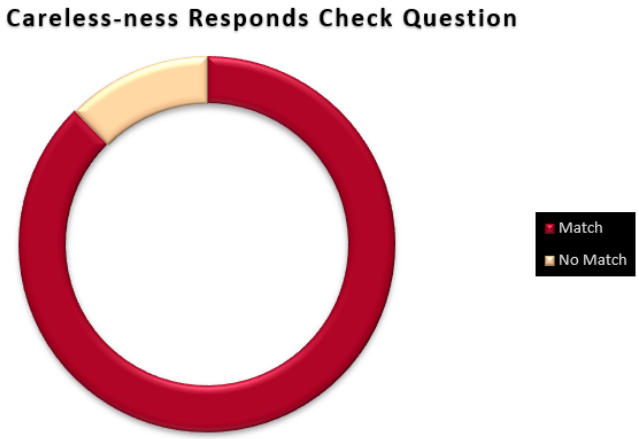
respondent's attentiveness. Additionally, response patterns and inconsistencies were examined to identify potential signs of carelessness in the data.

TECHNIQUES USED IN OUR SURVEY

We designed attention check questions which include questions that ensure the respondent's attentiveness and the integrity of their responses.

In Conclusion, We discovered that out of the 250 total respondents, 31 of the surveys were carelessly answered, translating to a total of 219 non-carelessness responses.

Comparison Between Col (C and Q)	Count of Comparison Between Col (C and Q)
Match	219
No Match	31
Grand Total	**250**



The Analysis Phase

Introduction

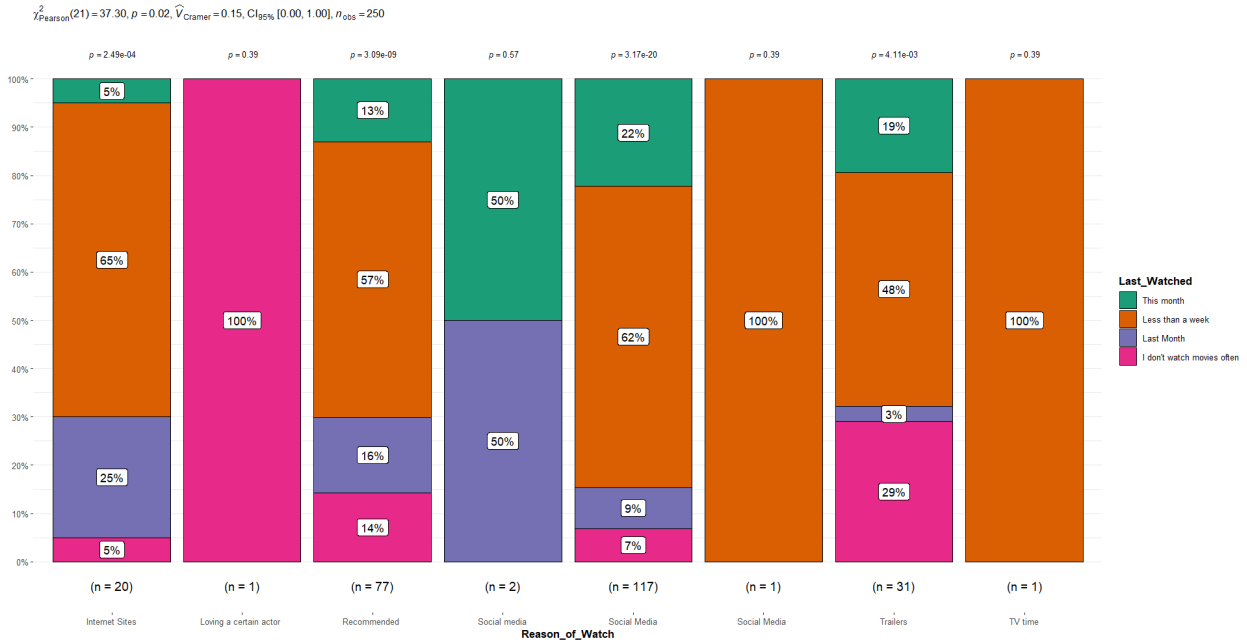
In this phase, we will go through the necessary statistics and analysis that will help us understand and predict what benefits our participants the most.

Descriptive Statistics

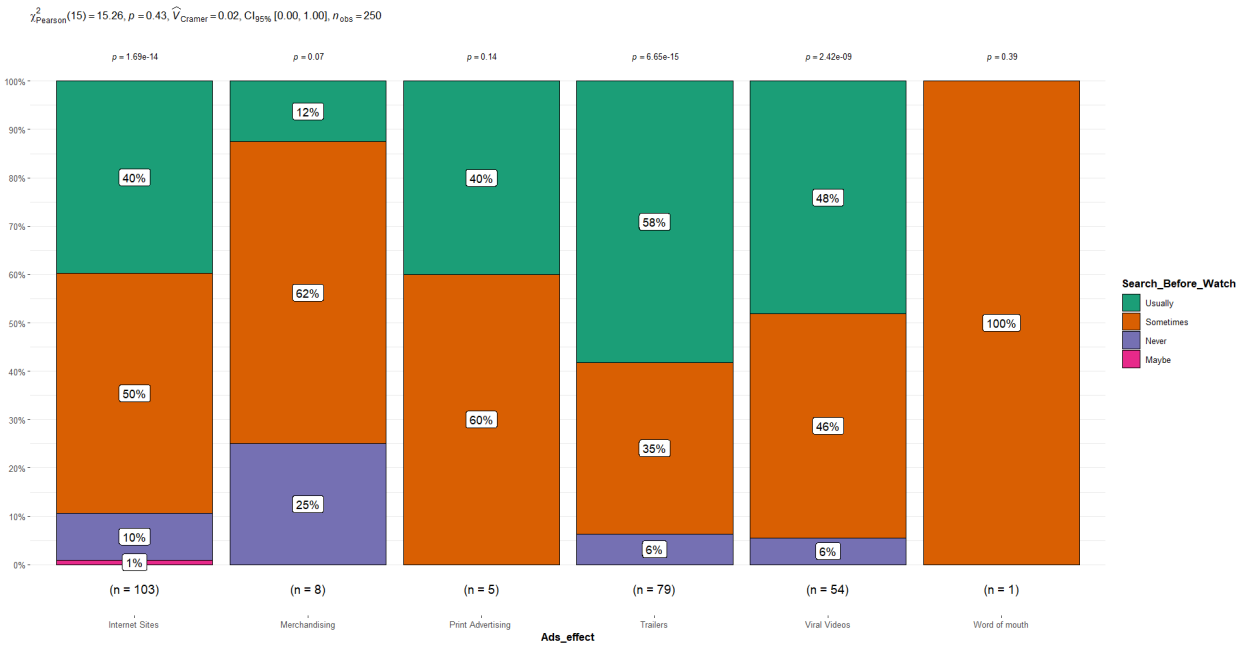
```
 34  Variables      250  Observations
---------------------------------------------------------------------------------------------------------------------------
Last_Watched
       n  missing distinct
     250        0        4

Value      I don't watch movies often                Last Month          Less than a week
Frequency                          30                        29                       147
Proportion                      0.120                     0.116                     0.588

Value                       This month
Frequency                           44
Proportion                       0.176
---------------------------------------------------------------------------------------------------------------------------
Frequent_Watch
       n  missing distinct
     250        0        4

Value            Less than once More than four times        Once or twice  Three or four times
Frequency                    53                   41                   87                   69
Proportion                0.212                0.164                0.348                0.276
---------------------------------------------------------------------------------------------------------------------------
Reason_of_Watch
       n  missing distinct
     250        0        8

lowest : Internet Sites          Loving a certain actor  Recommended            Social media            Social Media
highest: Social media            Social Media            Social Media            Trailers                TV time

Value              Internet Sites Loving a certain actor           Recommended          Social media
Frequency                      20                     1                    77                     2
Proportion                  0.080                 0.004                 0.308                 0.008

Value                Social Media           Social Media               Trailers               TV time
Frequency                     117                      1                     31                     1
Proportion                  0.468                  0.004                 0.124                 0.004
---------------------------------------------------------------------------------------------------------------------------
Frequent_Go_Cinema
       n  missing distinct
     250        0        4
```

Here we show the number of missing and distinct data values, and their frequency and proportion.

Also, we have done chi-square tests to show which columns are dependent and which aren't.

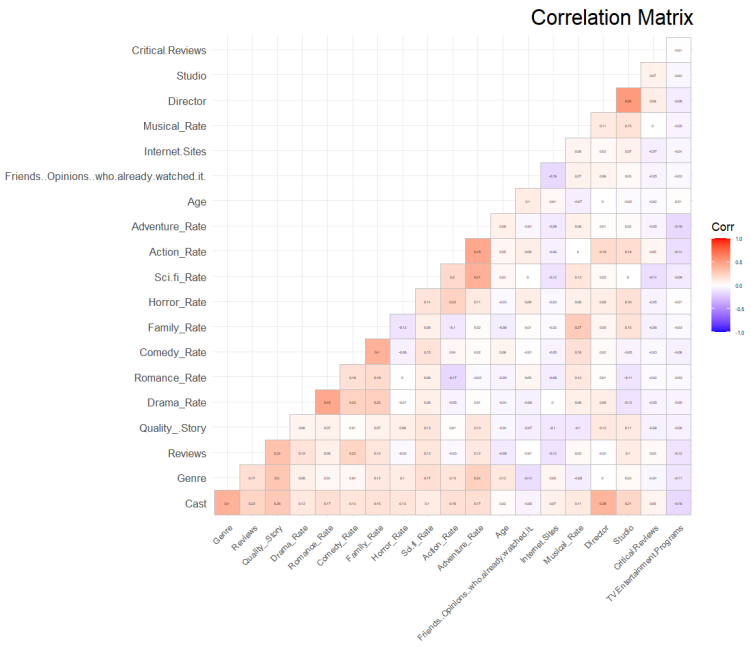


Here we conclude that those two columns are dependent on each other.



Here we conclude that those two columns are independent of each other.

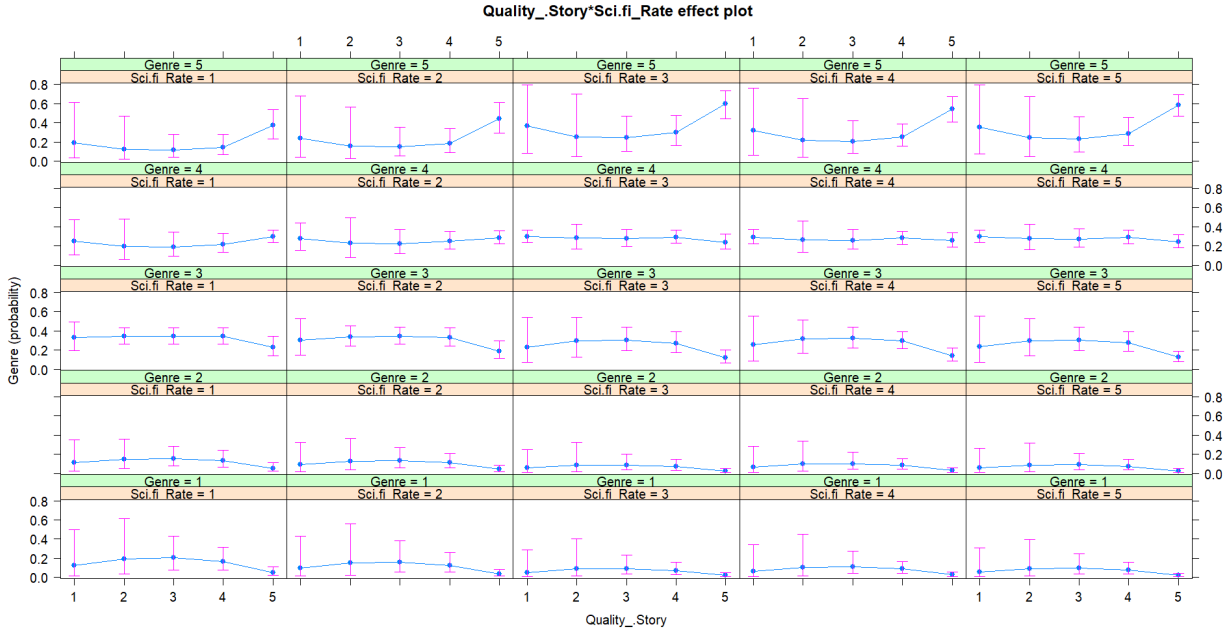
Correlation Task



Our correlation matrix here shows which values are dependent on each other.

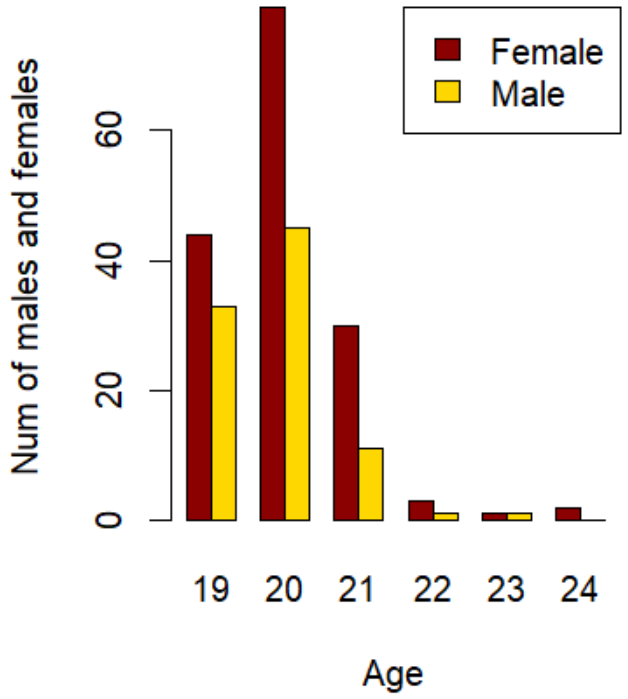
Regression Task

In this task, we will try to predict what our respondents like and dislike; to ensure a better experience for the audience.

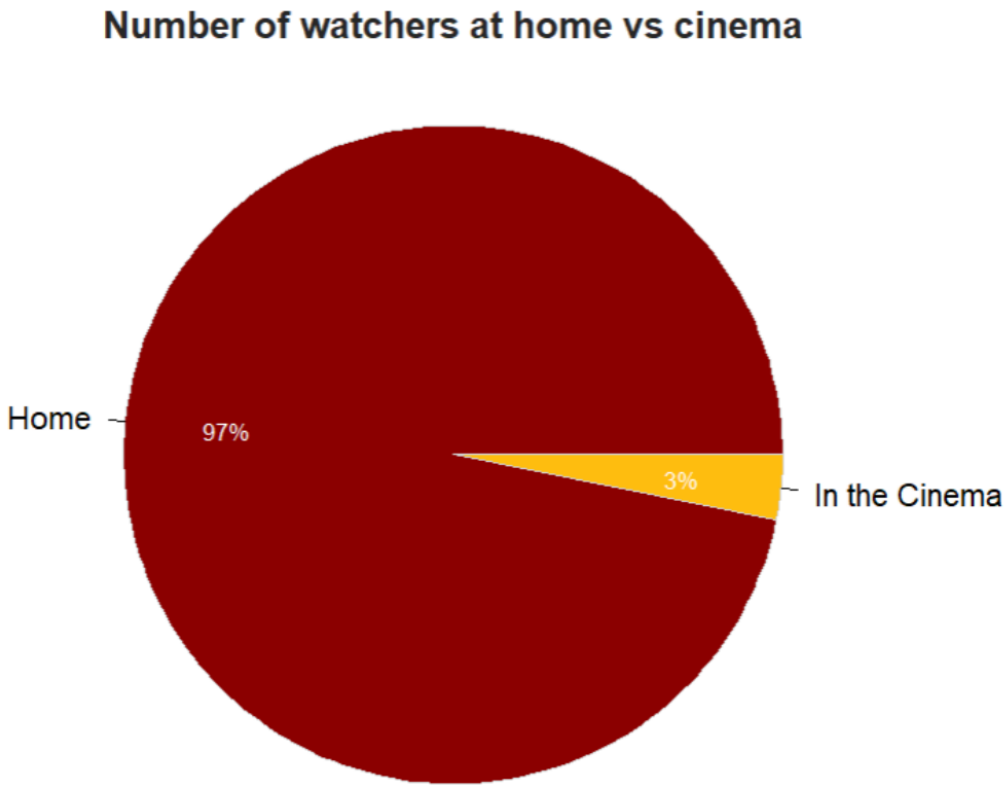


Here we see the effect of the independent sci-fi rate on genre and its probability.

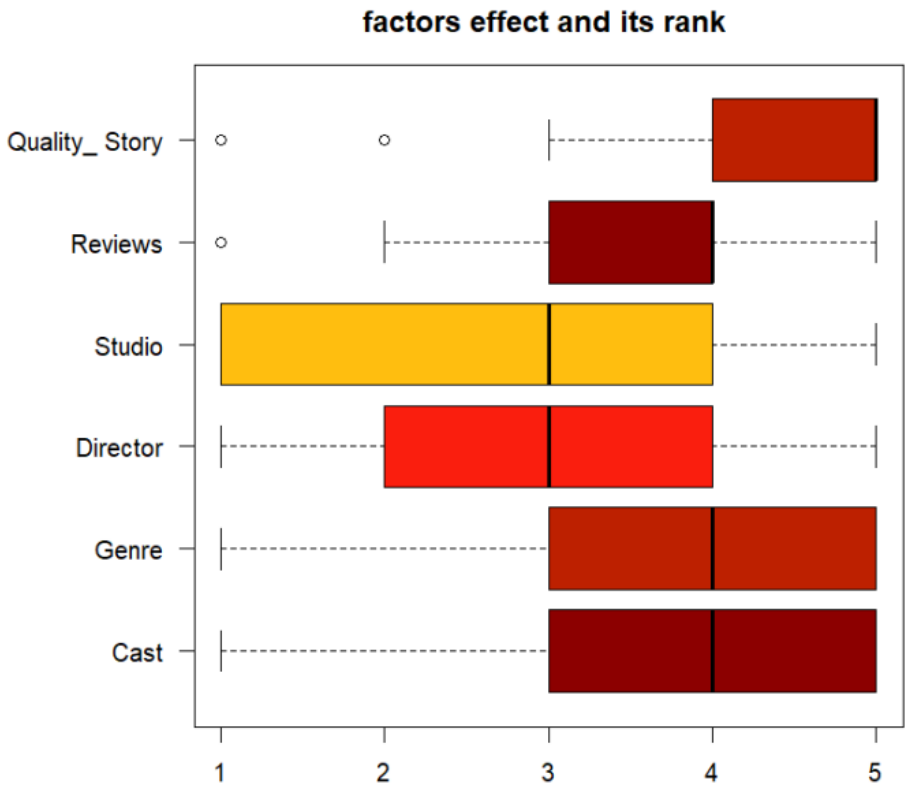
Data Visualization



We used multiple bar plots to know the number of females and their ages in this survey and males too. We conclude that the most frequent age that fills the questionnaire is 20 years and they are females.

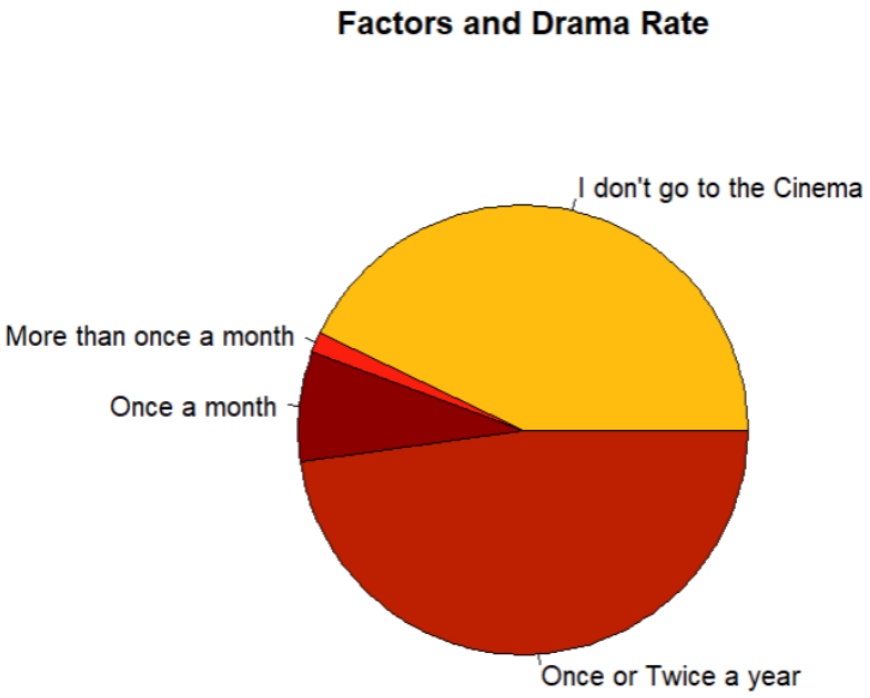


The number of people who watch movies at home is greatly higher than those who watch movies in the cinema.



This figure shows the distribution of every factor and its ranks.

- Quality of the story is very important for people.
- Reviews are important too but less than the quality of the story.
- Studio has most of the ranks (1,2,3,4) it differs from one to another.
- Director is normal not very important and not useless.
- Genre and cast are important for people (3:5).



Most people don’t go to the cinema or go once or twice a year.