Data analysis plan

Question 1: Is there a relationship between the type/genre and score?

Type:

To address this question, we have decided to make a hypothesis test. First of all, we should define our null and alternative hypothesis, which is that the overall score of Documentaries is higher than the overall score of Films:
H0: μD = μF (REWRITE)
H1: μD > μD (REWRITE)
The next step would be to determine the value of the sample statistic. We chose to use the DBM (Difference Between Means) because the size of the population is much larger than the sample size and samples are independent (or at least we consider so).
— · · —
xD - xF (REWRITE)
After applying the statistic of choice to the set of randomized samples, we would compute the p-value, which would tell us the proportion of samples on the distribution with an statistic as extreme or more than the observed sample.
If the p-value is lower than a 5% significance level (α=0.05 is used as a default level), the conclusion would be that the results are statistically relevant, thus rejecting H0. If the p-value is not lower than the significance level, it would mean that we cannot reject H0, and the results would be not statistically relevant.

Genre:

This question address a similar problem as the previous one, but with many different values for the qualitative explanatory variable. Despite not being sure on how to address this problem, our guess would be to perform the same analysis using an one-versus-all approach, computing the p-value of each of the different genres and comparing all of them individually against the significance level.
Additionally, we think that a Kruskal-Wallis one-way analysis would fit this case, indicating if wether at least one sample stochastically dominates the rest of the population or not.

Question 2: Is there any difference on the audience/critics/imdb score by genre?

To know which is the variable that has the most effect we could perform a Principal Component Analysis to measure the percentage of variance explained of each of the three variables (audience_score, critics_score and imdb_rating) to see which of them is tha highest. We should compare the three vectors of the n-dimensional data space, which we expect to be highly correlated.
Before doing so, we should standarize the explanatory variables to the same scale. We are not sure about how to do this, as the variables do not seem to be distributed equally between variables.
This would tell us which of the three variables should be taken into account the most.
As graphs suggest that critics tend to evaluate the lowest, we could perform an hypothesis test on each of the genres under the null and alternative hypothesis of the critics score being respectivelly equal or lower than the rest of the scores.
H0: μC = μO (REWRITE)

H1: μC > μO (REWRITE)

Question 3: Are oscar-awarded films more liked?

We find the procedure done at the first question to adapt well to this question too, so we would make the same test, with our null and alternative hypothesis:
H0: μO = μN (REWRITE)
H1: μO > μN (REWRITE)
Where the null hypothesis states that being nominated to the best picture award does not affect the overall score, while the alternative hypothesis states that this has a positive impact on it.

Question 4: What are the trends over the years?

Which is the referred month for releases? Over the years? Does it affect the score?

Despite seeing a clear trend of non-stationary data, after some research, we have concluded that the team lacks the knowledge to address (or even try to) this question yet.

Does the genre change over the years?

After some research, we have concluded that the team lacks the knowledge to address (or even try to) this question yet.

Do older films tend to have higher score/number of votes?

We would first perform a linear regression model to try to predict a film's score based on relevant variables, after normalizing them. Once we have done this, we can obtain the p-value of each of the variables based on the coefficients of the regression line. If the p-value of the date variable is lower than 0.05, we would consider this results statistically relevant and therefore conclude that there is a relationship (positive or negative, based on the coefficient) between the two variables (or so we think).

Question 5: Do actors and directors have a higher score once they won an Oscar?

This is a somewhat more complex question, as we are not well versed on the Time Series topic. However, we will try: our approach would be to divide the data for each ocurrence (directos and actors who have won an oscar) before and after winning the award. First, we should do some research to find the year they won the Oscar.
Once we have our data separated, we would perform a Time Series Regression for each of the ocurrences before and after. By comparing the regression slope of the variable score on the regression model, we could perform a hypothesis test to see if the change is relevant.

References: https://handbook-5-1.cochrane.org/chapter_9/9_2_types_of_data_and_effect_measures.htm
https://www.mathworks.com/discovery/time-series-regression.html
https://en.wikipedia.org/wiki/Stepwise_regression
https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance