**What one word describes how you instinctively feel about the work within the first 5-10 seconds? Is it positive or negative? – what are the good/bad thing about your dataset**

**Good qualities of the dataset**

**I would give it a 6 out of 10**

**Good things**

1.The structure in which the features of the single data items as well as the connections between them are shown must be used to represent datasets.

In order for automated analysis to be viable, a dataset format that is suitable for retention must maintain the syntactical consistency of the framework and individual values. Understanding the semantic of the data items and their interactions within the dataset is also crucial for future usefulness. The semantics may be directly stated within the dataset, openly stated in a supporting document (ideally one that is also machine-processable), or implicitly stated by adherence to a standard operating procedure or external specification.

2.Depending on the intended usage and domain-specific needs, a dataset format may accept few or many different data types for values or attributes. The commonly used CSV (comma-separated values) form, which lacks explicit data type and is typically restricted to text (ASCII or Unicode) representations of text and integers, represents one extreme of data representation.

3.Providing assistance for particular software interfaces

According to the definition of scope above, features that are important to the current generation of users who share their characteristics or to a larger group of users in relevant disciplines may also be important for long-term preservation. The capability of a form to facilitate effective analysis of a kind suited for a field or data category is one particular factor. Many datasets, for instance, ought to be presented in a way that permits common statistical analyses like inter, t-tests, regression analysis, or principal component analyses. A table or collection of tables can be used to represent data in a number of different data formats.

Bad things

1.The data was not normalized i.e. the id of the movies

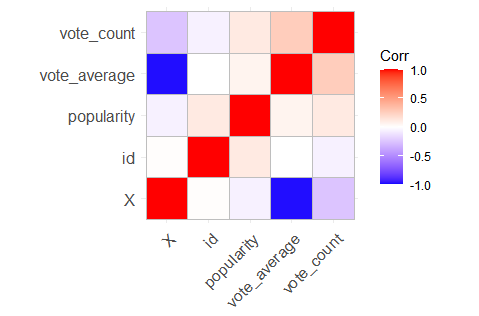
2.Some data was in categorical format which made it difficult to perform descriptive statistics .

2.The date format also had an issue.

Question 2

Very subjective but do you like the visualisation (might be the subject or visual form)? What score on a scale of 0 to 10 would you give it (10 is best)? Consider what factors influenced your ratings? Rate it based on how much its easy or hard to learn And how much you think it can be utilized in your profession

**Using a heat map**

****

I’d give it a 9 rating

The importance of a heat map

How it can be used in my profession

In case of how movies are rated either by average view time , number of clicks and basically customer behaviour. Many of these questions can be answered with heatmaps. Data visualization has been around for many years, and it always helps to view information differently. Heatmaps for websites are used to understand data that would otherwise be presented in a cumbersome spreadsheet.

Imagine them as an extra layer added on top on your website that captures a variety of user actions, such as scrolling, clicking, and overall grabbing their attention, while yet protecting their privacy, which is extremely important in today's society.

The trend of movie profits over time for various genres is seen in the plot below. It seems that recent years have seen an increase in the profitability of "Family" and "Adventure" films, while "Drama" films have been making less money.

Question 3

Do you feel the project successfully – and sufficiently – facilitates understanding (does it help you learn something about the subject matter or, at least, confirm/reinforce what you already knew)? What score on a scale of 0 to 10 would you give it (10 is best)? Consider what factors influenced your ratings?

The first project was more difficult than any of our previous quizzes. I think the quizzes were beneficial in helping me get ready for project 1 and feel at ease with it. Project 1 forced me, in my opinion, to apply my analytical thinking and problem-solving abilities in a constrained

length of time. I gained more knowledge about statistical coding because to the project, which is

vital whenever it comes to advancing in a Data Scientist profession. I'd say go for it.

an 9 because statistical coding was not as heavily emphasized in earlier instruction as it was in the first project

4. Consider the project’s effectiveness or otherwise in demonstrating the principles of trustworthy, accessible and elegant design: where does it succeed and where does it fail?

**Tustworthy data**

Instead of using absolute values, utilize relative values for comparing across locations, categories, or groups.

**Accessible**

• Is targeted to the audience;

concentrates on the pertinent information;

is neither overly simple nor overly complex; considers potential disabilities (such as color blindness);

and is suitable for the format presented.

**Elegant**

According to shaker , Don't create anything until it is both essential and beneficial, if it is both, go ahead and create something lovely.

What would you have done different

**Lack of enough elegance**

Also learning to create more elegance visualizations.I feel that the color selection was a bit off , giving a 5 out of 10

Question 5

.Whilst you may not know much about the project’s hidden context, what would you do differently? How would you help to get these pair of ratings higher towards the maximum of 10?

I would have spent more time getting to know the given data sets and all of their potential uses. Before going to the residency, I would have spent more time learning more about statistical coding so that I could have explained and understood the codes' outcomes better. This would have made it easier for me to modify the code so that it better suited my data collection.

Chapter 2

**DEVELOPING INTIMACY WITH YOUR DATA**

i) Our dataset

Using the str (mydata) function i was able to view the data structure of my dataset .It had 9481 observables and 8 variables.

>

| str(mydata) 'data.frame': 9481 obs. of 8 variables: |
| --- |

ii) The dataset had 8 columns 5 numerical 3 categorical.

The following are ways for univariate data analysis

1. Summary statistic -Determines the value's center and spread using summary statistics.

2. Frequency table - This displays the frequency at which different values occur.

3. Charts: A graphic display of the data distribution.

Chapter 2

1. Normalization -The data was stored between the values o and 1. Each value is divided by the maximum value of the feature in the "basic feature scaling" method. Consequently, the new values fall between 0 and 1.
2. Eliminating some columns- we eliminated the review columns which had very long texts describing the users view on the individual movies
3. Dealing with missing values/blanks

There were no missing values in our data