CHAPTER 4
ANALYSIS AND DESIGN

4.1 Analysis

This Chapter explains regarding problems and formulation of design to be completed with an in depth analysis, thus it is presented in details as clearly as possible, a narration, diagram is presented in this chapter and a picture to clarify the narrative is supplied.

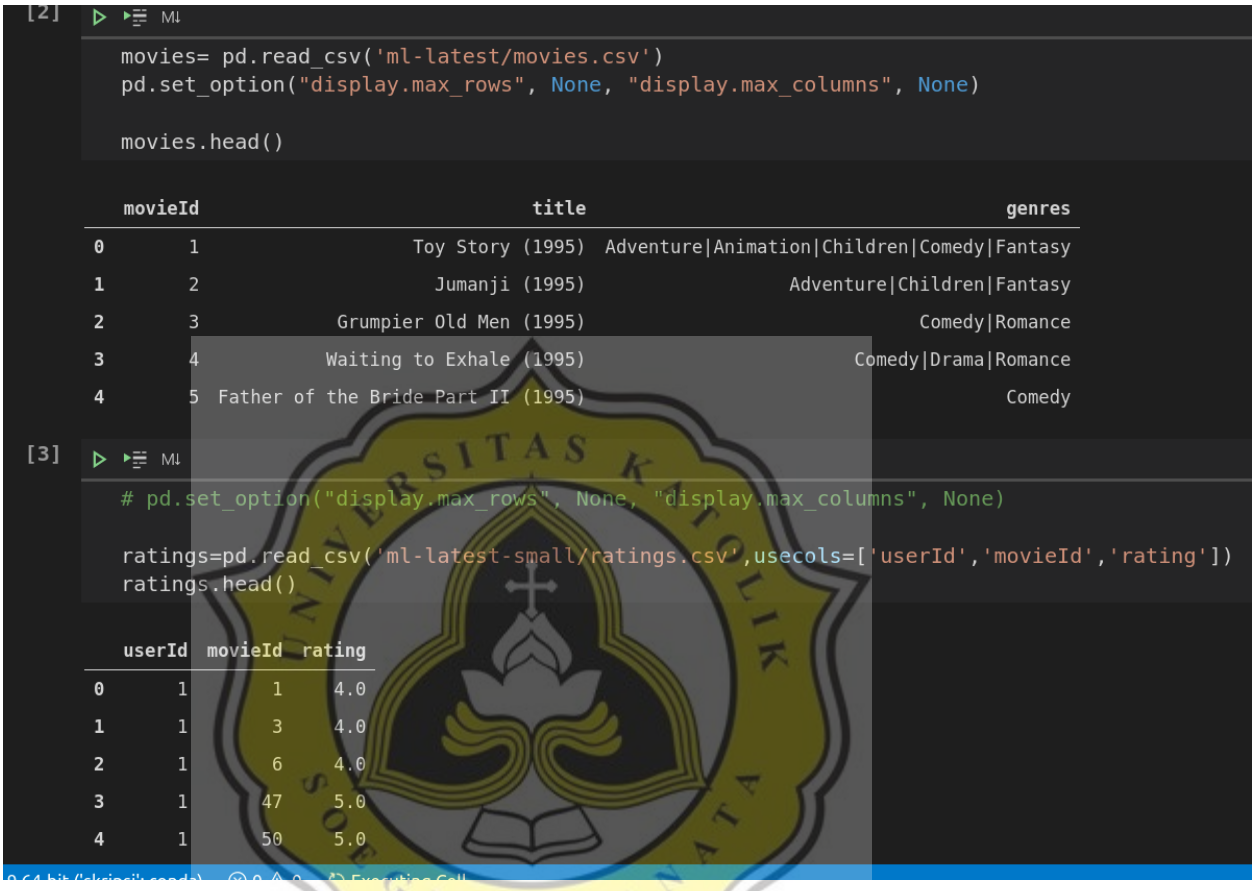
4.1.1 Collecting Data

Data collection usually require several phases to be able to complete, in this case data is obtained from grouplens.org a website that host movie rating and corresponding user in raw format or linked format. The data collected is stored in multiple csv files to be accessed using pandas Dataframe. The main dataset that will be used is as follows but not limited to 'ratings.csv', 'movies.csv', 'tags.csv'.

4.1.2 Data Pre processing

Data pre processing is a process in which we look for all possibilities inside the dataset. In the commonly distributed database system, the database administrator will query all table, join, trigger and something related to find as much information as possible to help make the data easier to read, easy to understand structure and better improvisation for the goal of the project which is to come up with a goal for what the data will be utilized for according to its original planning such as ranking system. In Recommender system data preprocessing help us understand how the data behave in such a way that implementing a function and algorithm will make more sense than ever before, in short data preprocessing is the way we query all the csv files description(), info() and stats of the particular data row such as data type (int32, float64, object, etc.). Merger of dataset is needed in this project. The merged dataset is the movie dataset and the rating dataset, both of them

contains userID in order to merge them into group for the purpose of comparison of the dataset between ratings dataset and movies dataset.



[2]

```
movies= pd.read_csv('ml-latest/movies.csv')
pd.set_option("display.max_rows", None, "display.max_columns", None)

movies.head()
```

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

[3]

```
# pd.set_option("display.max_rows", None, "display.max_columns", None)

ratings=pd.read_csv('ml-latest-small/ratings.csv',usecols=['userId','movieId','rating'])
ratings.head()
```

	userId	movieId	rating
0	1	1	4.0
1	1	3	4.0
2	1	6	4.0
3	1	47	5.0
4	1	50	5.0

Figure 4.1.1 Data Pre-Processing or Querying

Data collected for querying can be represented in a bar graph, pie diagram, or many plotting system such as matplotlib to visualize the data in the right way based on how the data will be analyzed and viewed for the goal of the project or solution, this will most likely vary depending on the need of analytics such as sparse graph with random data point usually used in population analytics or other statistical method or KNN Model that needs a lot of data point, thus a normal pie chart or bar diagram may not fit the needed visualization of the dataset. Plotting is done by calling the matplotlib library and filling the desired output to the parameter such as pie chart with percentage point.

4.2 Images and Flowchart Design

This chapter will present Design types and diagram for the purpose of this project making it understandable in the implementation section later in Chapter 5, images, charts / flowcharts and diagram is presented in this section.

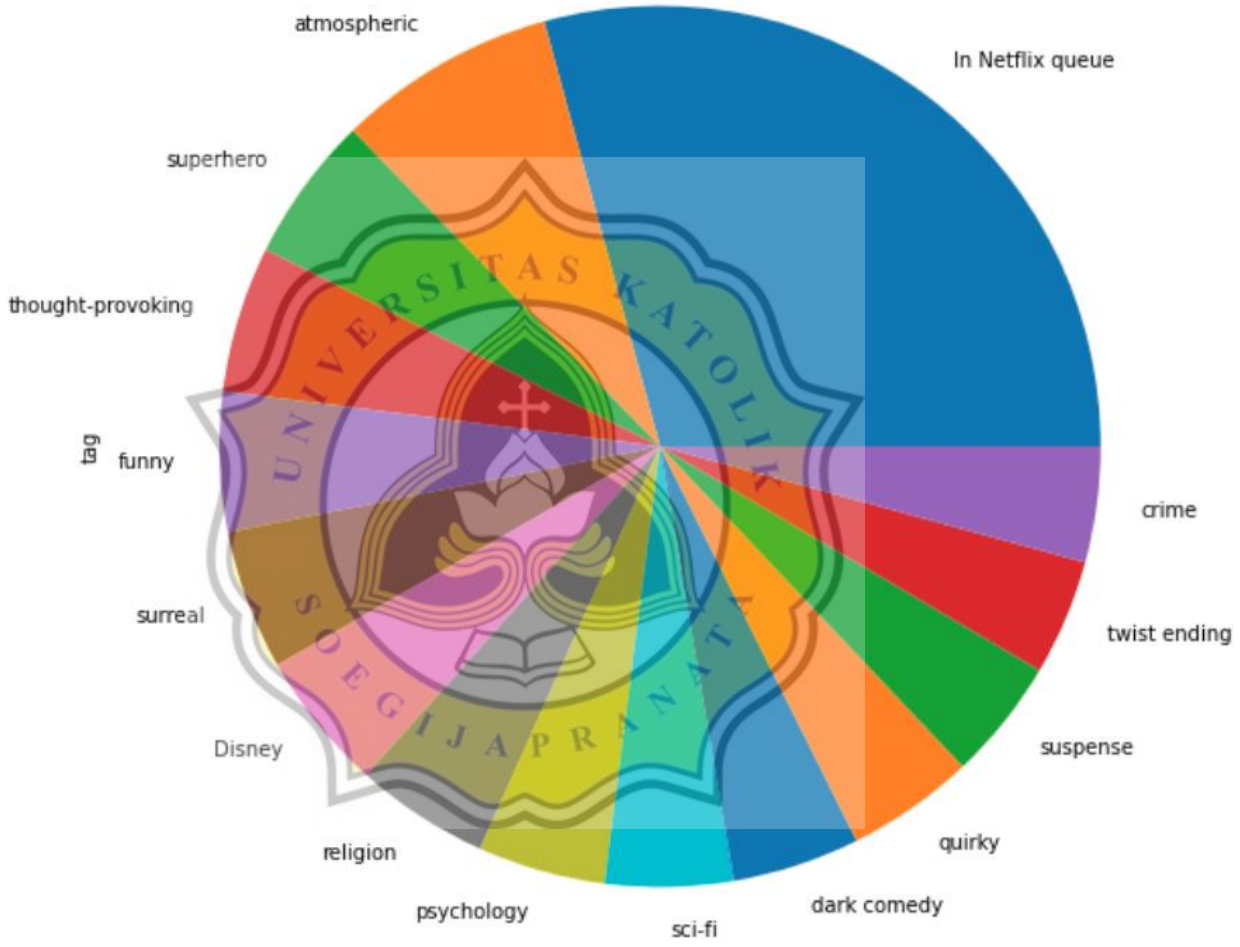
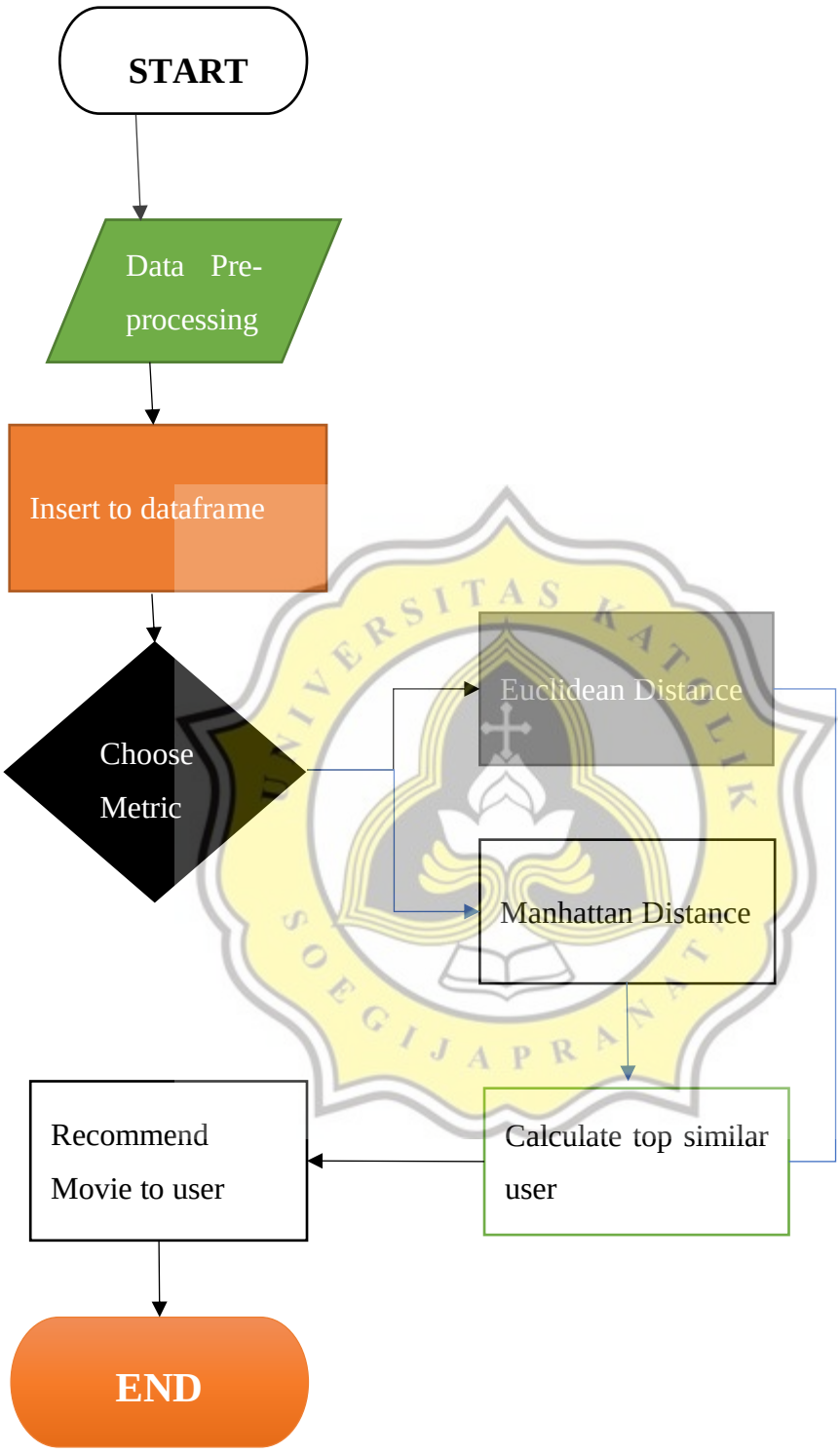


Figure 4.2 Movie Category Chart

The Plotted image can tell us how the dataset behaves instead of just showing the first 10 or 20 lines of a data frame here we can see how the data looked like in a bigger picture sorted neatly in a categorical order. For example the purple colored partition on the right shows 4.2% of the movie is tagged as visually appealing by the user who rated that particular movie.

In a Recommender System normally there is a table that shows movie name and a genre on the top rows to confirm the property of the movie. This table by definition is actually a grouped values that aggregates the individual items of a more extensive table within one or more discrete categories. The summary can include sums, average, or more statistics.


START
Data Pre-processing
Insert to dataframe
Choose Metric
Euclidean Distance
Manhattan Distance
Calculate top similar user
Recommend Movie to user
END

4.3 Tabel

This table shows the difference between Manhattan distance and Euclidean distance of 2 user's with the rating dataset of about 100,836 record. The narrowed version of about less than 200 record for rating will be presented in the next chapter at the testing section of the page.

Table 4.3.1 Distance comparison

Comparison of distance metric		
User ID	***Euclidean Distance***	***Manhattan Distance***
User 1 & User 9	0.5	0.5
User 1 & User 21	0.011940298507462687	0.13793103448275862
User 1 & User 310	0.06060606060606061	0.1
User 11 & User 41	0.023121387283236993	0.06060606060606061
User 61 & User 89	0.13793103448275862	0.2857142857142857
User 23 & User 86	1. 0	1.0
User 1 & User 85	1.0	1.0
User 1 & User 77	1.0	1.0
User 1 & User 53	0.5	0.5
User 1 & User 44	0.06666666666666667	0.09090909090909091

4.4 Function

A taxicab geometry or in this case called Manhattan distance is a type of geometry in which the traditional distance function or metric of Euclidean geometry is substituted by a new metric in which the distance between two points is equal to the sum of their Cartesian coordinate absolute differences.

$$d1(\mathrm{p},\mathrm{q}) = \left|\left|p-q\right|\right| i = \sum_{i=1}^{n} |pi - qi| \quad (1)$$

Figure 4.3 Manhattan Distance

This formula explains the sum of p subtracted by q in an absolute value meaning that all the value will not be a negative value since the purpose of this calculation is to see how far the distance of user's movie taste between one user to the other targeted user. Later in the code the value

will be returned so that a value between 0 and 1 is achieved and no value outside the range is allowed for easier readability and comparison with other distance metric i.e., Euclidean distance score.

.

$$d(p, q) = \sqrt{(p - q)^2} \quad (2)$$

Figure 4.5 Formula Euclidean Distance

This formula explains how Euclidean distance is calculated. Firstly it counts the inner part of square root (p-q) squared, then a square root is applied and a returned result will appear, in the case of user's preferences similarity the value will be adjusted to allow a range of 0 and 1 for easier readability and comparison with Manhattan distance.