|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ujian Akhir Praktikum - Semester Ganjil 2020/2021**  *Practicum Final Exam –Odd Semester Year 2020/2021* | | | | | | |
| **Matakuliah**  *Subject* | | | **COMP6579 – Big Data Processing** | | |  |
| **Kelas**  *Class* | **:** | **BB08** | | **Tanggal Mulai**  *Start Date* | **: 19 December 2020** |
| **Waktu Mulai**  *Start Time* | **: 11:20** |
| **Dosen**  *Lecturer* | **:** | **D6198 - Fepri Putra Panghurian, S.Kom., M.T.I.** | | **Tanggal Selesai**  *End Date* | **: 19 December 2020** |
| **Waktu Selesai**  *End Time* | **: 13:20** |

**PERATURAN UJIAN:**

*Exam Regulations:*

* Mahasiswa tidak diperbolehkan berdiskusi dan/atau bekerja sama dengan peserta ujian lainnya

*Student is not allowed to discuss and/or work together with other exam participants*

* Mahasiswa tidak diperbolehkan untuk membuka dan menyalin dari **BUKU** atau **CATATAN**, **VIDEO** dari pengajar (recording kelas, VBL, Youtube, dsb) dan **REFERENSI** lainnya

*Student isn't allowed to open and copy from any resources such as notes, videos (class recording, VBL, Youtube, etc) and other references*

* Mahasiswa tidak diperbolehkan membuka dan menyalin jawaban dari internet (google, stackoverflow, dsb)

*Student isn't allowed to open and copy answer from the internet (google, stackoverflow, etc)*

* Asisten **BERHAK** memberi nilai 0 **(NOL)** bagi peserta ujian yang melakukan segala bentuk kecurangan

*Assistant is able to give 0 (ZERO) score for exam participant who does any cheating actions*

* Kumpulkan jawaban tepat pada waktunya, apabila terlambat mengumpulkan maka jawaban tidak akan dikoreksi dan nilai mahasiswa adalah 0

*Submit the answer on time, if not, then the answer will not be checked, and the students will receive 0 (ZERO)*

* Bila Anda tidak membaca peraturan ini, maka Anda dianggap telah membaca dan menyetujuinya

*If you have missed to read these regulations, so you are considered to have read and agreed on it*



**SOFTWARE YANG DIGUNAKAN:**

*Software will be used:*

* VM Cloudera
* Jupyter Notebook

**FILE YANG DIKUMPULKAN:**

*File must be collected:*

* IPYNB

**PERHATIAN!**

*Attention!*

* Bagi yang mengerjakan tidak sesuai dengan soal, maka akan diberikan nilai **NOL (0)**

*For those who do not work in accordance with the exam case will be marked as* ***ZERO (0)***

* Bagi yang mengerjakan tidak sesuai dengan software dan versi yang telah ditetapkan, maka akan tetap dikoreksi dengan software dan versi yang telah ditetapkan

*For those who do not work in accordance with the software and specific version will be corrected by the predefined software and version*

* Kompres semua jawaban yang akan diunggah. Pastikan format pengumpulan nama file dan ekstensi sesuai dengan format berikut: **[NIM]-[NAMA].zip**

*Compress all file that will be uploaded. Make sure the format for collecting file name and extension according to the following format:* ***[NIM]-[NAME].zip***

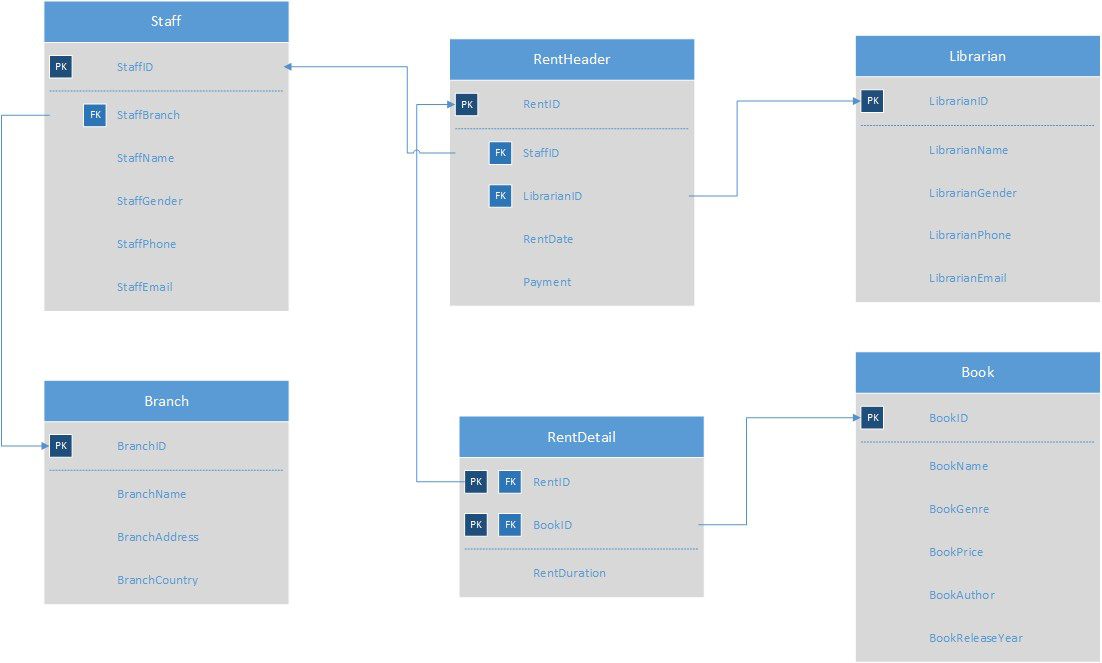
## Soal

*Case*

**Grimoire Library**

**Grimoire Library** is an awesome store to buy or rent books. Since the sales are going high, they need to improve the store sales more quickly. To do that they intended to do some **analysis** with **different** **kinds** of data they have.

From the sales business process, the data can be analyzed to gain sales insight. The data is stored in **Comma-Separated** **Values (CSV)** file and the data schema is drawn using **Entity Relationship Diagram (ERD)** below:



**Figure 1. Grimoire Library ERD**

Below is the task you must do to analyze the data:

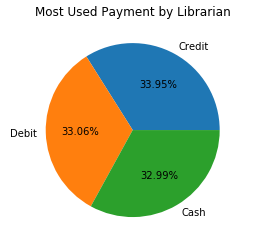
1. **Load Data from CSV to Spark**

Using **SparkSession**, **read** the following files (“Librarian.csv”, “Branch.csv”, “RentHeader.csv”, “RentDetail.csv”, “Staff.csv”, “Book.csv”).

1. **Query Analysis and Visualization**

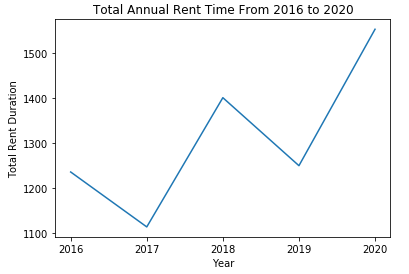
You are asked to gain some sales insight about the data. Below are some statements you need to answer. Use **SparkSQL** to answer the question and **pyplot** package to **visualize** the answer.

1. Show the **percentage of payment used** for paymentthat has **at least been used 1000 times and above** using **pie plot**.



**Figure 2. Most Used Payment Statistics Figure**

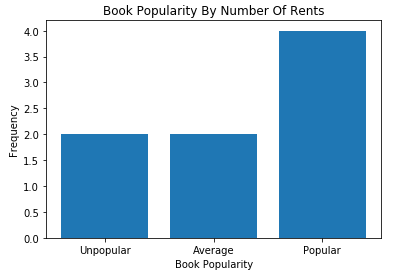
1. Show the **total** **annual rent time** from **year 2016 until year 2020** using **line plot**.



**Figure 3. Total Annual Rent Time Statistics Figure**

1. Show the **frequency** **of book genre’s popularity** where the book genre is **any, but horror and thriller are excluded** using **bar plot**. The **book genre’s popularity** if the **total number of rents** reached the following condition:

|  |  |
| --- | --- |
| Popularity | Number of Rents |
| Popular | >= 600 |
| Average | 575 - 599 |
| Unpopular | < 575 |



**Figure 4. Book Genre’s Popularity Frequency Figure**

**Healthy LyfeStyle**

You are scientist researching on how human daily activities affects their health in order to know the ideal healthy lifestyle is. Based on the data you have collected, you decided to make a **clustering** model that will divide your data into **2 cluster**, **healthy lifestyle cluster** and **non-healthy lifestyle cluster**. You will be given **“Healthylyfe\_Train.csv”** and **“Healthylyfe\_Test.csv”**, and here is the description of columns:

|  |  |
| --- | --- |
| Column Name | Description |
| Name | The person’s name. |
| Diet | The person’s junk food consumption in a month. |
| Exercise | All person’s exercise hour in a week |
| Marrital Status | Whether the person is married or not (Married,  Not Married). |
| Body Height | The person’s body height in centimeter. |
| Sleep Quality | The person’s quality of sleep (Poor, Average,  Good). |
| Age | The person’s age in year. |
| Sick | How often the person gets sick (Rarely, Sometimes,  Often). |

**Figure 5. Healthylyfe\_Train.csv**

|  |  |
| --- | --- |
| Column Name | Description |
| Name | The person’s name. |
| Diet | The person’s junk food consumption in a month. |
| Exercise | The person’s exercise hour in a week. |
| Marrital Status | Whether the person is married or not (Married,  Not Married). |
| Body Height | The person’s body height in centimeter. |
| Sleep Quality | The person’s quality of sleep (Poor, Average,  Good). |
| Age | The person’s age in year. |
| Sick | How often the person gets sick (Rarely, Sometimes,  Often). |
| Healthy | Whether the person has a healthy lifestyle or not  (Yes, No). |

**Figure 6. Healthylyfe\_Test.csv**

Below are the steps you are required to do to generate the model:

1. **Load Data**

Given the file **“Healthylyfe\_Train.csv”** and **“Healthylyfe\_Test.csv”**, you are asked to load the data using **SparkSession**.

1. **Select Features**

After you load the data, you need to **select important features** that will be used for training. Pick **three important features.**

1. **Data Preprocessing**

In this step, please remove any **missing values** in the data.

1. **Transform Data**

In this step, transform the raw data so that it is suitable for training. For example, **recode** the ‘**Marrital Status**’ column value to be either **0** or **1**.

1. **Normalization**

After data preprocessing, you are required to **normalize** the data. Use the **StandardScaler**

package to normalize the data.

1. **Generate Model**

Next, you are required to **generate** a **model** from the data. Use the **KMeans** package to generate the model into **2 cluster**.

1. **Visualization**

After the model is generated, you can **visualize** the model using the **pyplot** package. Don’t forget to add **x-label**, **y-label**, and **title** for your plot.

1. **Model Testing and Evaluation**

Then, you can **test** the model to check predict whether the data will be in **healthy lifestyle cluster** or in the **non-healthy lifestyle cluster**. Print the accuracy of your model and get the **model** with **minimum accuracy 80% or higher**.

**Please ask your teaching assistant if there are any related questions.**

**Good Luck 😊**