**CS 486/586 Introduction to Databases Project - Part II**

**Database Project Implementation:**

●      **Part 1: Project proposal (Due date - 3rd week, Oct 14th) - 20 points**

●      **Part 2: Mid project (Due date - 7th week, Nov 10th) - 200 points**

●      **Part 3: Progress report (Due date - 8th week, Nov 18th) - 30 points**

●      **Part 4: Final project (Due date - Finals week, Dec 7th) - 150 points**

●      [5 points] Show that you have successfully installed Postgres and you can use pgAdmin or psql on your local machine or virtual machine (you can provide a screenshot).

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

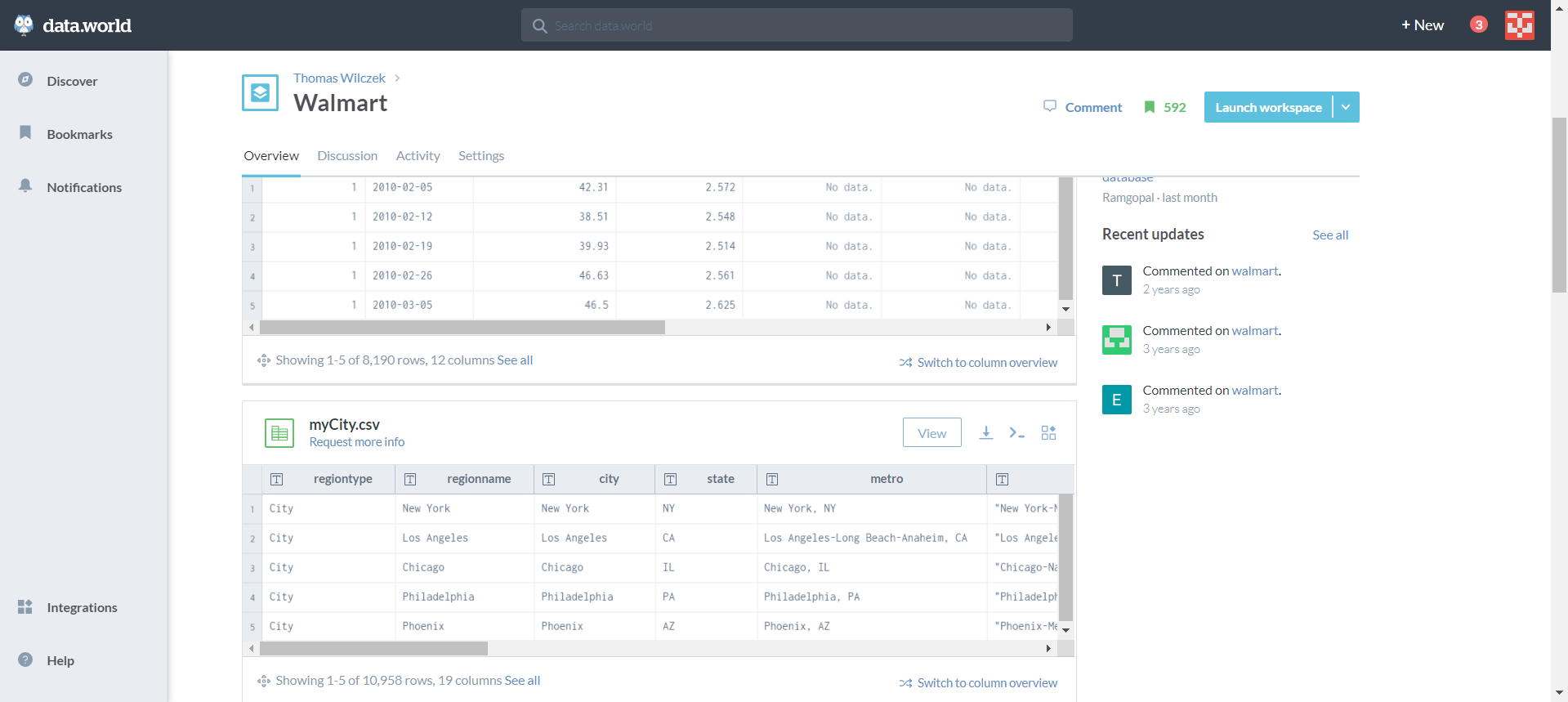
●      [20 points] ER diagram.

●      [25 points] Database schema designed with your designed tables, attribute declaration, primary/foreign keys, views, or temporary tables etc. (You should demonstrate a variety of schema and SQL features such as a variety of data types, keys, foreign keys, different cardinalities).

●      [45 points] Data loading process and data preprocessing and cleaning. You should have effective data entry and avoid large amounts of manual data entry. Please describe what you have done to clean your data fully, what you have considered in cleaning, and how you have chosen to load and import your data in your tables.

**Data Processing**

**Collecting data:** Initially Data is collected from website called data.world: <https://data.world/tommywilczek/walmart> the data is in the form of structured but the data is not-normalized and also the data is having noise data(noise data: Data which is inaccurate and disturb the results Ex: data type is numerical but data present is string, which will cause the problem during the processing)



**Data Cleaning:**It is also known as scrubbing. This task involves filling of missing values, smoothing, or removing noisy data and outliers along with resolving inconsistencies

**Data cleaned steps for this data:**

1. **Mismatched data types-**Checking the data type within each column
2. **Missing data-**Initially checking with null values

**Noisy data**

Data cleaning also includes fixing “noisy” data. This is data that includes unnecessary data points, irrelevant data, and data that’s more difficult to group together.

Removing the unwanted data or replacing the unwanted data

Google Co-lab link : <https://colab.research.google.com/drive/1QyN5RnQB6m5MAUHtXOTOe5r3Zbbc4L10?usp=sharing>

Importing libraries for processing

Graphical user interface

Description automatically generated with low confidence

Reading data From CSV to Data Frame

Graphical user interface

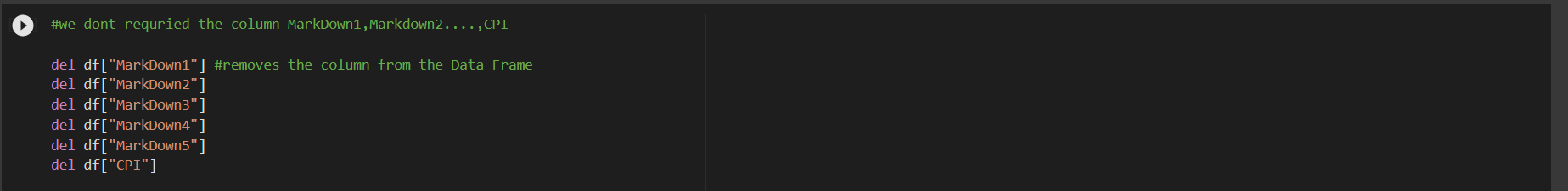
Description automatically generated

Checking the columns

Background pattern

Description automatically generated

Removing Unwanted Columns

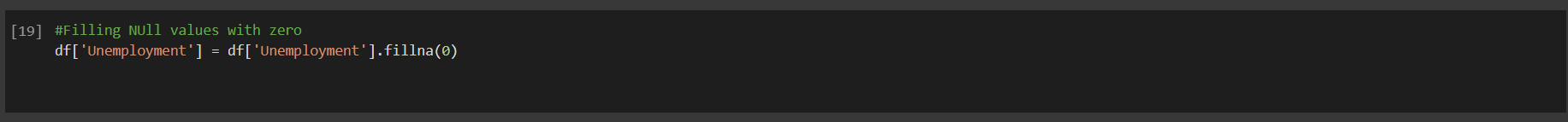


Finding null values

Background pattern

Description automatically generated

Filling Null values with Zero



After Filing with zero

Graphical user interface, application, Teams

Description automatically generated

Checking the column data type

Background pattern

Description automatically generated

**Normalizing Data**

**Unemployment data TABLE**

**Table

Description automatically generated**

**Week\_sales table**

**Table

Description automatically generated**

As we can see the holidays is common for every store at particular date so we can just create a new table for the holiday, so that we can easily access the data from new table created.

**ISHOLIDAY TABLE**

**A picture containing text, crossword puzzle

Description automatically generated**New Table for the Isholiday

**Pulling data from stocks**

Want to add another table with Walmart stock market values at particular dates, for that I’ve used the python and yfinance(Yahoo Finance) library