### **2.Data Description**

#### ****2.1 Reprocess the Original Dataset from internet****

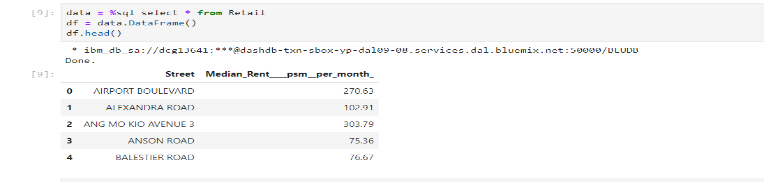
#### I obtained the Singapore 2020 Q1 retail rental dataset from Singapore Government official website. <https://www.ura.gov.sg/-/media/Corporate/Property/Commercial-Properties-2020Q1/comm_median_rentals_2020Q1.pdf?la=en>. The Dataset contains the information of Rental rates of different Singapore Streets per square feet for retails and offices in Q1 2020. Below graph is an example of the dataset in pdf format.

#### 

#### As mentioned above, the original file is a pdf format file. I tried to using Tabula.read\_pdf function to extract the table from the original pdf file but failed to do so in the end. Thus, I downloaded the data directly from the link and converted it to CSV file in my computer. I did the data cleaning for the missing values and the unwanted columns in the CSV file by using excel.

#### 

#### After data cleaning, the CSV file was uploaded to IBM Db2 on cloud. Use SQL magic retrieve the data and convert to dataframe in the Jupyter notebook. With this dataframe, we could predict locations with ideal rental rate and analysis further by using Foursquar API.



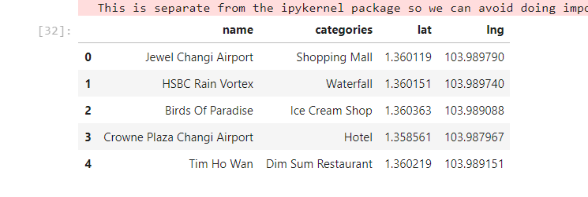
**2.2 Retrieve all the Latitudes and longitudes for all the streets in the dataframe**

By installation of Geopy package, all the location details for the streets in the dataframe could be found out. We could retrieve the Latitudes and Longitudes as below:



**2.3 Get Near-by-venues occurrence frequency for all the streets in the dataframe**

After getting all the latitudes and longitudes, it was ready to connect with Foursquare API to explore the near-by-venues for all the streets. The Foursquare Credentials and version were defined before I explored. For the first street I explored, I got the top 100 venues that are in the street within a radius of 500 meters. I borrowed the ***get\_category\_type*** function from the Foursquare lab to clean the json and structure it into a pandas dataframe.



I repeated the above process for all the streets in the dataframe. and grouped the rows by neighborhood (street name) and by taking the mean of the frequency of occurrence of each category. Now it was ready to analysis the near-by-venues to determine the suitable locations to start a new spa business: any bus stop or MRT station there? Any spa centres already opened there? Is there near to the shopping mall or hotel?

