ADTA 5240.403 Final Project

Suhas Siddarajgari Tellatakula 11626111

Part 1 – Big Data Strategy

Need for Big Data:

Our company has various such solutions/services in business which have the potential to grow in the near future. With the current company's minimal location-based infrastructure, it is highly difficult and impossible to scale these solutions to serve a huge bandwidth of customers irrespective of location. GCP has exact storage services to serve this purpose. Based on my understanding of our firm's data and keeping our domain specific guidelines and security regulations, GCP can provide authorization, role-based access etc. In order to tackle the huge competition with our industry's leading performers, big data technology offers data analysis providing insights and trends which play a huge role in taking critical decisions keeping our company far ahead of our competitors.

With the integration of Cloud services, we can increase the memory space and control any load Using as many servers as needed. We can pay and get these services only for what we used. And we can access these services through internet connected devices from anywhere. These offer many advantages over the existing infrastructure like higher Productivity, remote working, easy Collaboration, Redundant Backups, Reliability, scalability and High Security.

Why GCP?

I propose ***Google Cloud Platform*** services over other cloud service providers as GCP is simple to set up and configure initially and gives quicker responses at better prices. Also, GCP provides wide range of services like Compute Engine, Databases/Storage Migration, Developer

Tools, Data Analytics, AI/Machine Learning Management Tools, Internet of Things, Serverless Computing, Networking, and Operations.

Implementation Strategy:

As the Chief Data Officer of this company, I propose the following strategies to incorporate Big Data into our company's technology infrastructure or ecosystem serving the entire Data Analytics lifecycle from data preparation, preprocessing, migration, evaluation and visualization:

a. ***Compute Engine***:
   In order to overcome many restrictions and drawbacks with the traditional setup computing infrastructure, we can use GCP's Compute Engine which basically lends their infrastructure to us, and we can utilize as much as we need, and we pay only for the infrastructure and for the time we used. It is basically a virtual machine of customized CPU, GPU and other hardware like network devices through which we can perform our daily operations logging in from anywhere. This eliminates most of the on-site issues with IT setup, power, back up and so on. Because of its flexibility, we can scale up for large scale data processing or down our infrastructure needs when the resources are not much necessary.

b. ***DataProc:***
   Inorder to incorporate the Apache Hadoop Spark for distributed processing, we can use GCP's DataProc. DataProc is fully managed, easy to use service where we can employ Hadoop spark clusters of sizes depending on our data. It is simple and easy to administer. Open-source tools can also be easily integrated.

c. ***Cloud Storage:***
   In order to store all kinds of structure, unstructured, semi structured data our company generates, gets, logs, shares can be stored with GCP's cloud storage which is incredibly advantageous and

profitable with features like easy access, security, scalability, performance and many more. We can include Apache Hadoop ecosystem as our file management system as well.

d. ***Cloud SQL & Cloud Spanner:***
   As our company needs database services for our everyday operations, GCP's database services like Cloud SQL which is a most used relational database for data creation and configuration. Cloud SQL is managed entirely by Google saving us the maintenance, management, administration and so on. Like Cloud SQL, Google cloud also offers Cloud Spanner which is highly consistent and better availability with OLTP Online Transaction Processing kind of database which can be scaled for further purposes.

e. ***Big Query:***
   We can use Big Query in GCP for our Data Analysis solutions as Big Query can provide insights on given data irrespective of it's size, be it in Terabytes with web interface. It is the best query processing fully managed service for our company's data analytics.

f. ***Open Refine***:
   Like solutions such as Microsoft Excel, Open refine can be used for advanced data processing. It is perfect for any kind of processing tasks like data cleaning, data wrangling, data duplication, merging data and so on.

g. ***Looker Studio***:
   Google's Looker or Data Studio is another very useful service for Data Visualization. We can easily generate different kinds of charts, graphs etc., and can share them easily across our company. This helps in finding data insights helping reach correct business decisions faster.

h. ***IAM & KMS***:
   Since we are deploying every minute details of our company's precious data with someone outside, security is the biggest concern one can have. This issue is resolved with GCP's Identity Access

Management and Cloud Key Management Services. KMS can generate and destroy cryptography encoded keys providing secure user access and compliance. Also, IAM gives us the option to maintain who can access what in our company's data.

i. ***Cloud Shell/Google Cloud App***:
Finally, in order to access all these different services at one place, we can download the Google Cloud Application or use it through command line using Cloud Shell.

Steps to implement technology:

1. First and foremost, we must create an awareness in the company about this technology shift or upgrade.
2. Then, we hire a senior level Big Data specialist or expert to execute this transformation in systematic way.
3. Next, we must train our technical employees on above GCP/ Big Data tools and services through hands-on training sessions, learning tutorials etc.
4. We implement this technology upgrade through a pilot project in one least affecting department and perform all kinds of testing and get feedback.
5. Based on the feedback, implement the suggestions and employ the same in all other departments in a phase wise manner.
6. Finally, create a small team to constantly help and monitor these services and continuously improve the system.

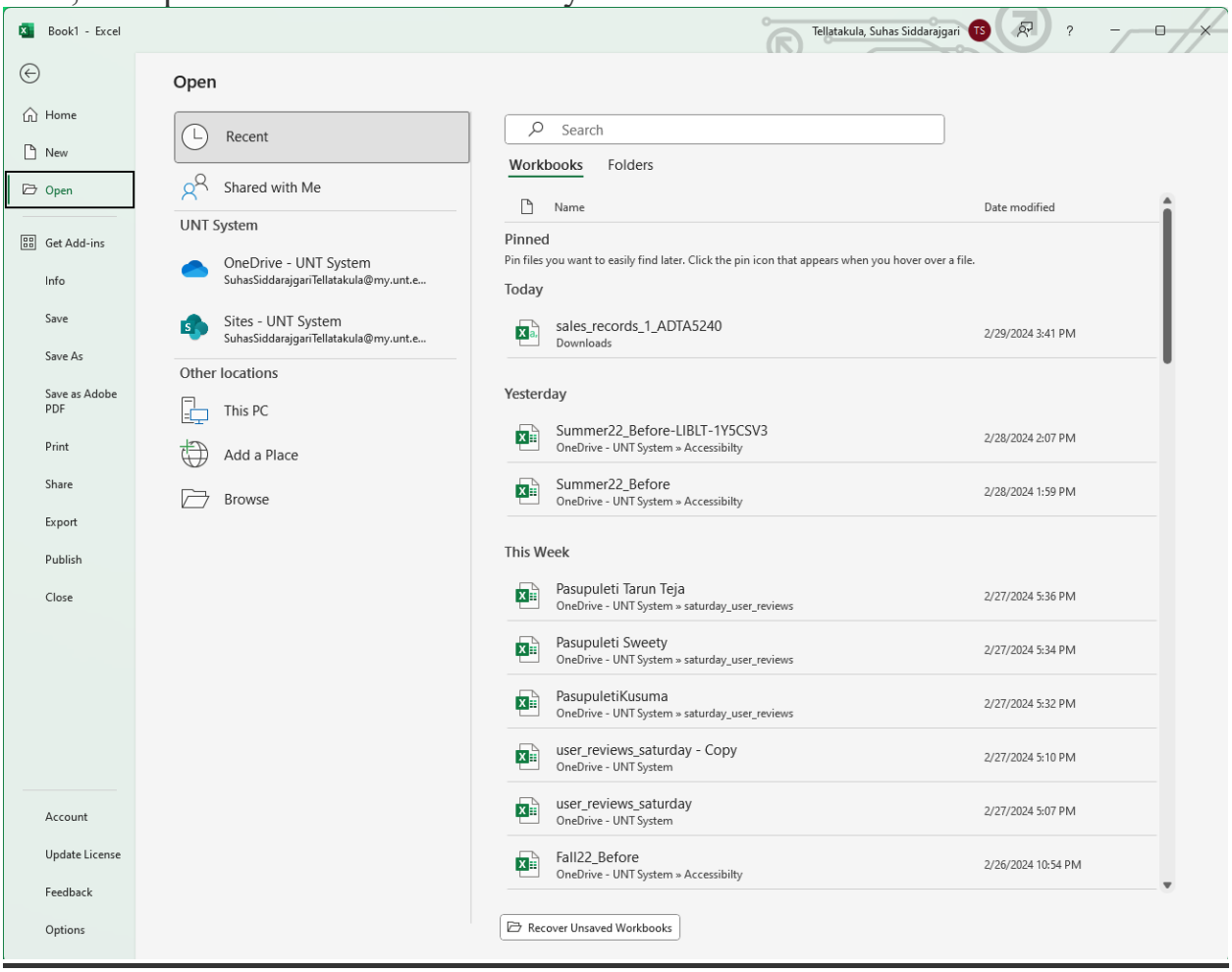
Conclusion:

As the technology industry is rapidly developing, bringing us new possibilities, if we do not make use of such latest technology at a right time, we may go behind in the tech business. With the above suggested strategy, we can take advantage of the latest Big Data technology before our competitors and boost our company's efficiency and performance greatly in a brief time in a profitable way. This technological adaptation will place us ahead in our industry.

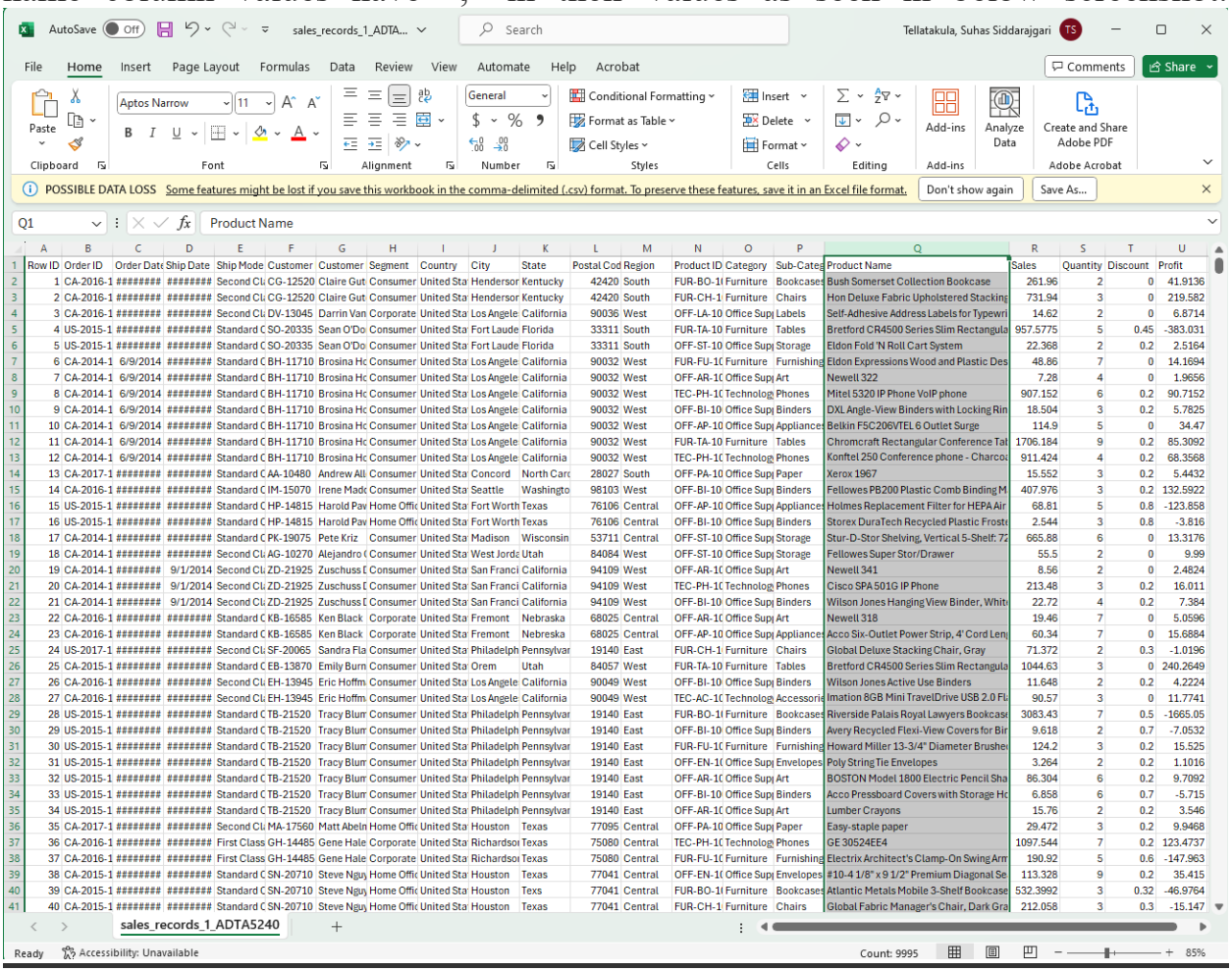
Part 2a – Data Preprocessing

For this part, I prefer using Microsoft excel for data preprocessing. First I downloaded the **sales_record_1_ADTA5240.csv** from datasets module in canvas.

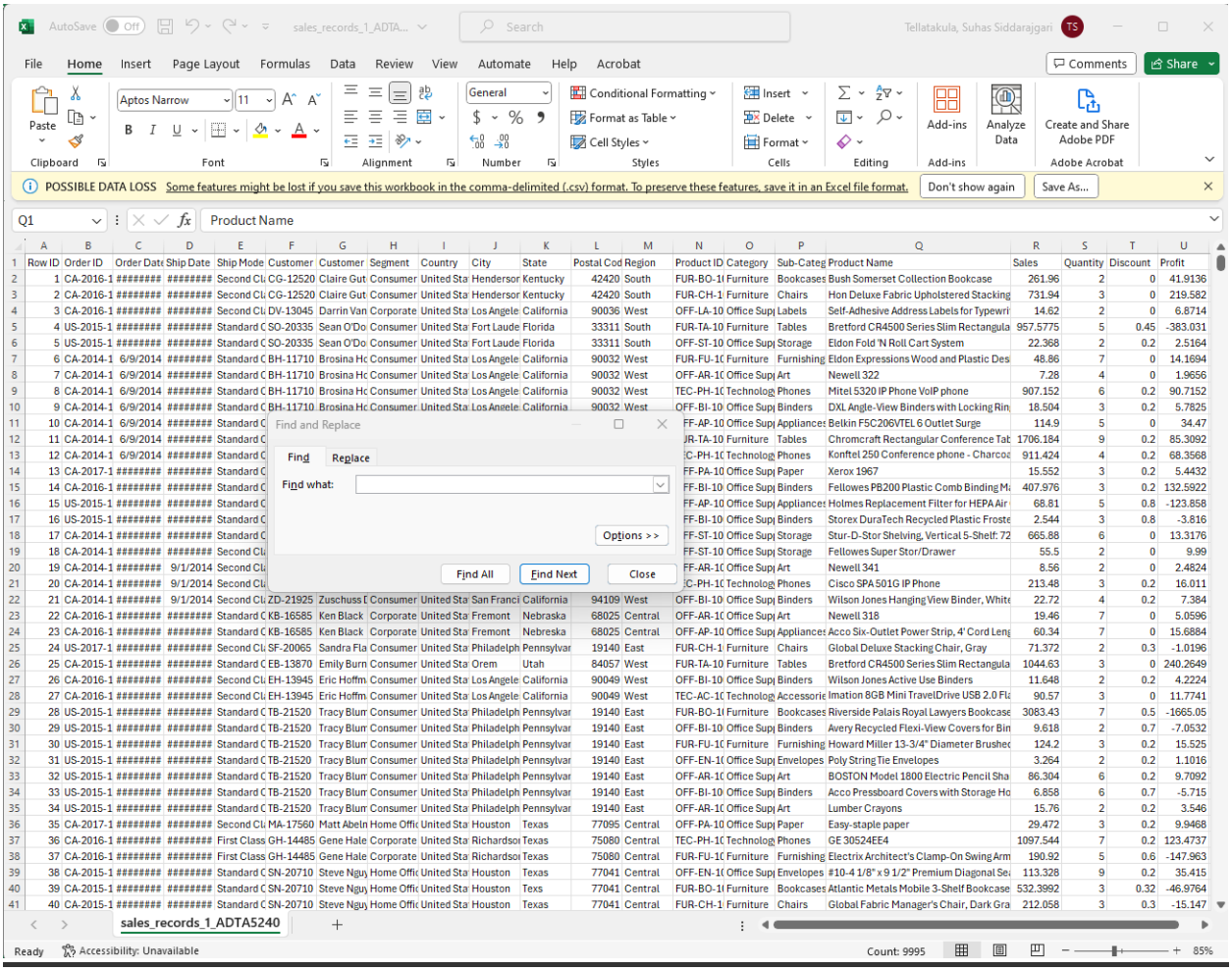
First, I imported this dataset into my excel as shown in the below screenshot.



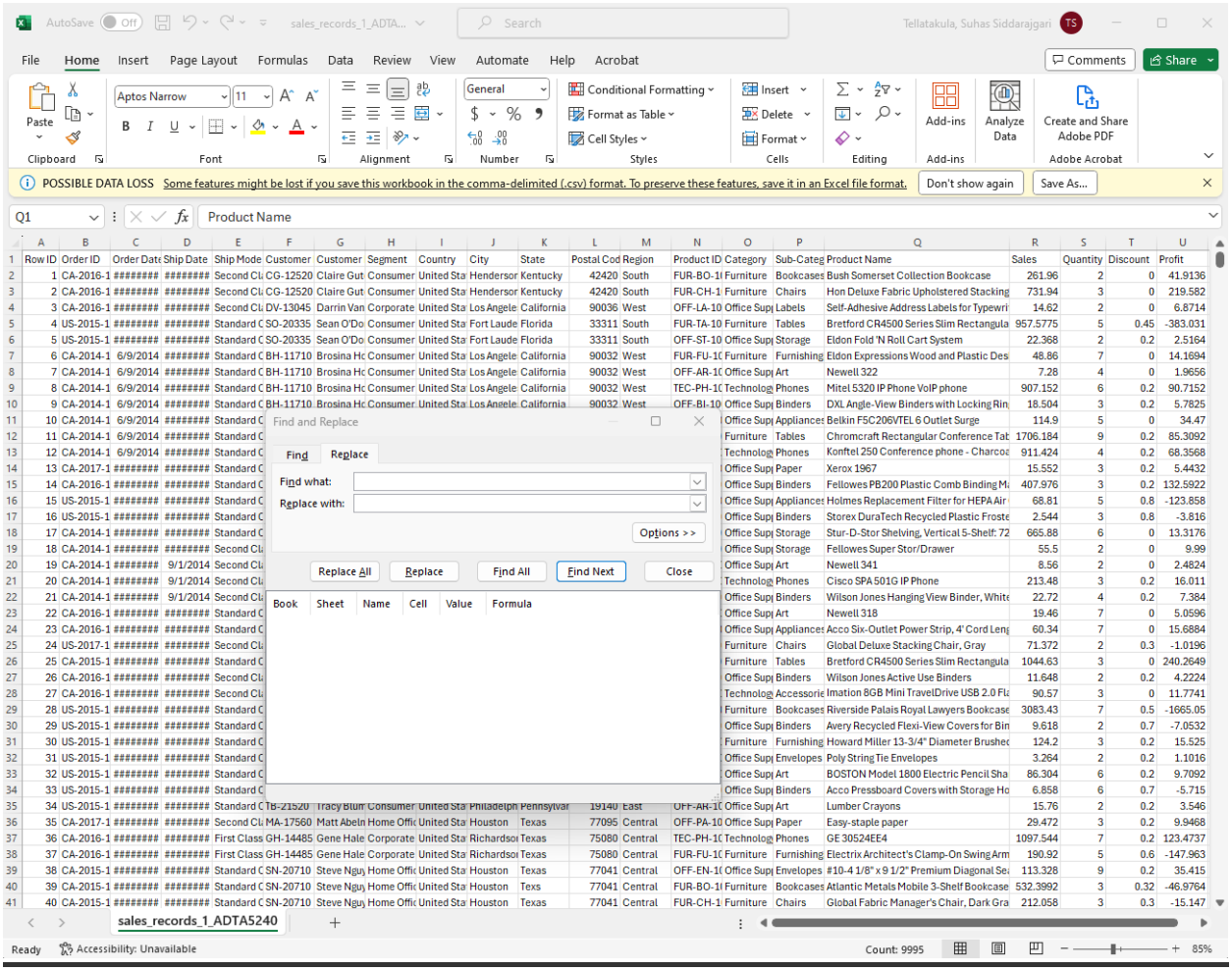
Once it is imported, we can see the data inside the sales records with various details about orders, customers, products, states, quality etc., We can see that the Product name column values have “,” in their values as seen in below screenshot.



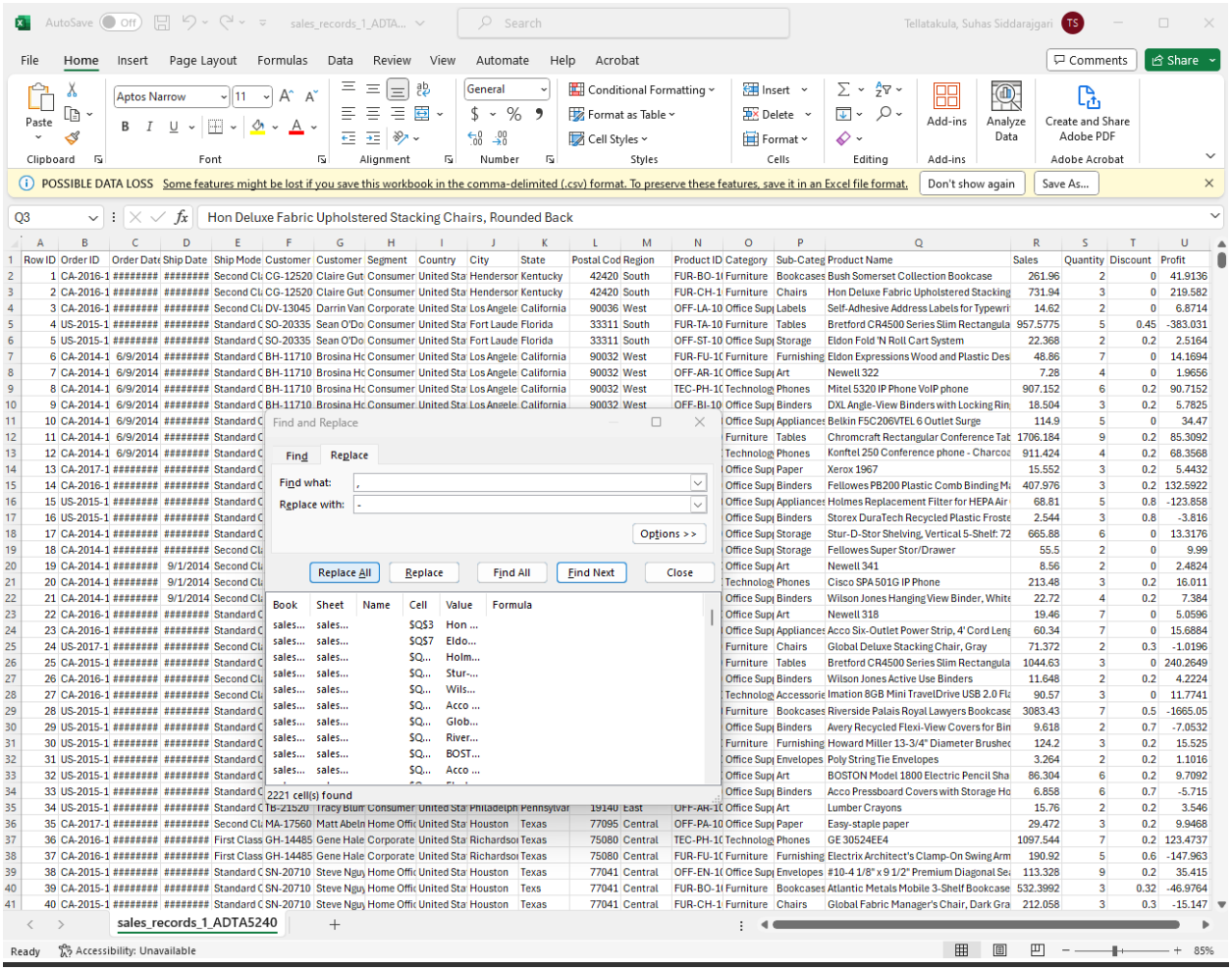
Next, I use a shortcut Ctrl+H for “Find and Replace” feature in Excel selecting the Product Name column which pops up a form like this with find and replace tabs as shown in below screenshot.



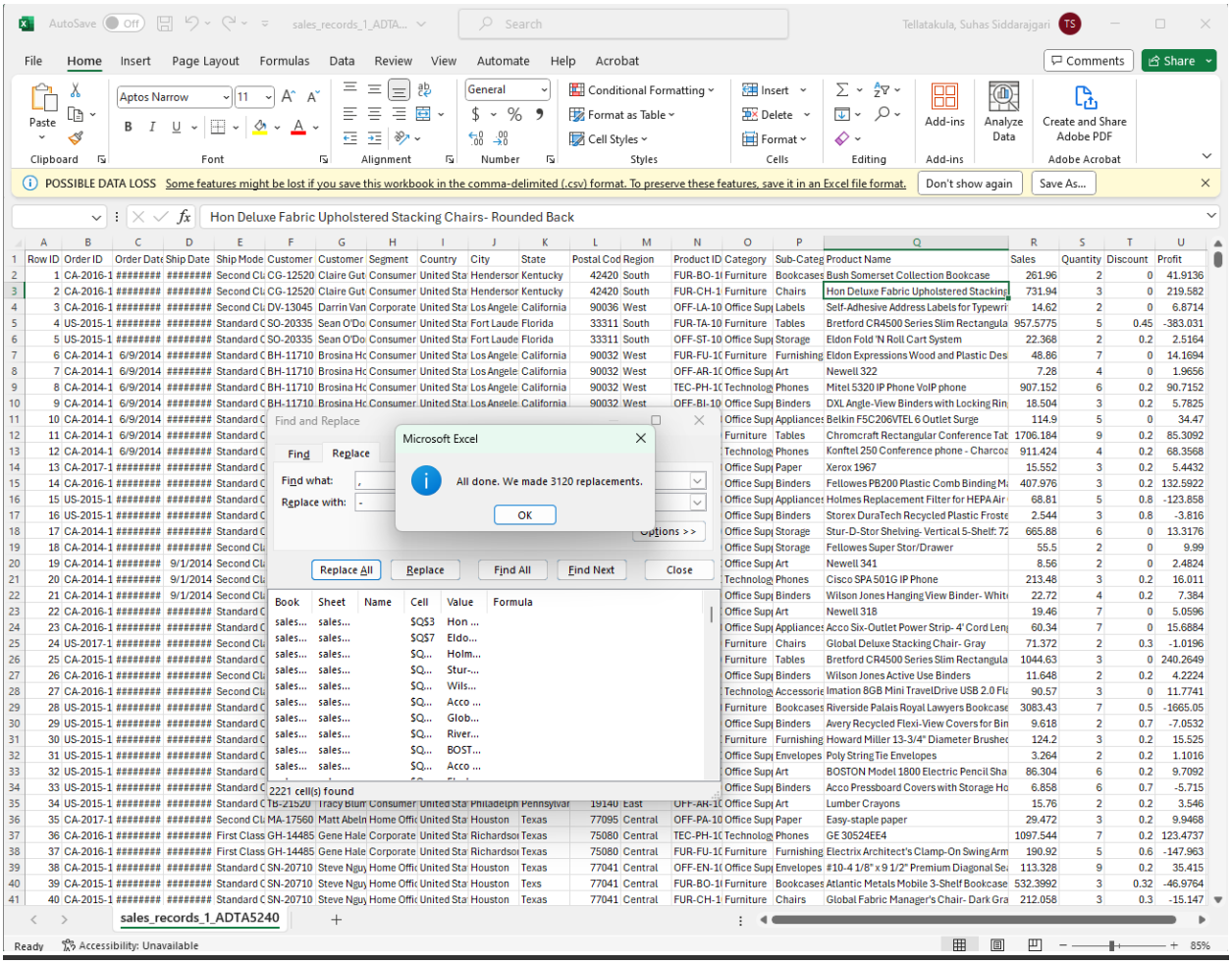
When clicked on the Replace tab, we can give what we are finding and what we wish to replace it with as shown in below screenshot.



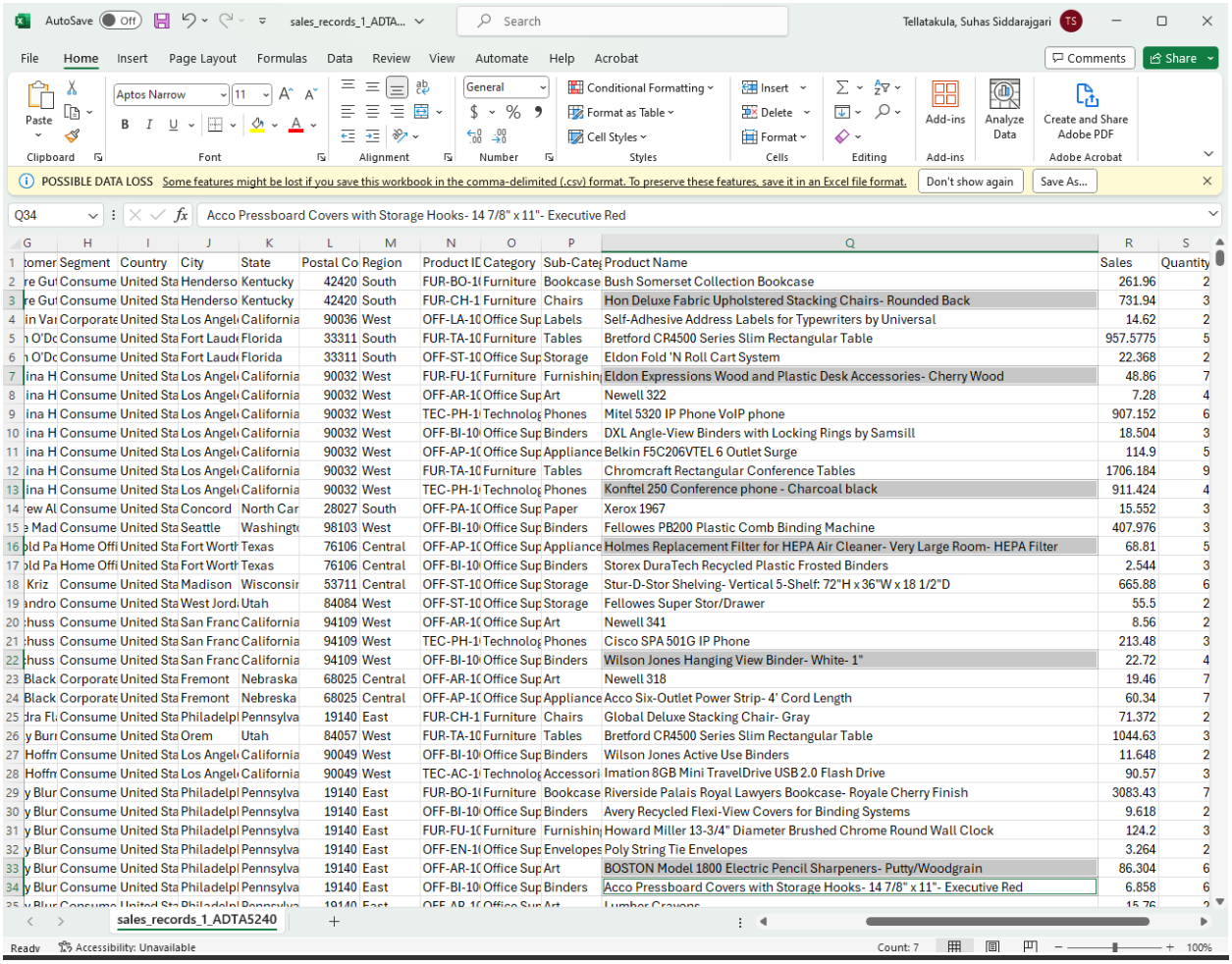
As we are looking to replace the commas with hyphens, we mention those values and click on the Replace All button as shown in below screenshot.



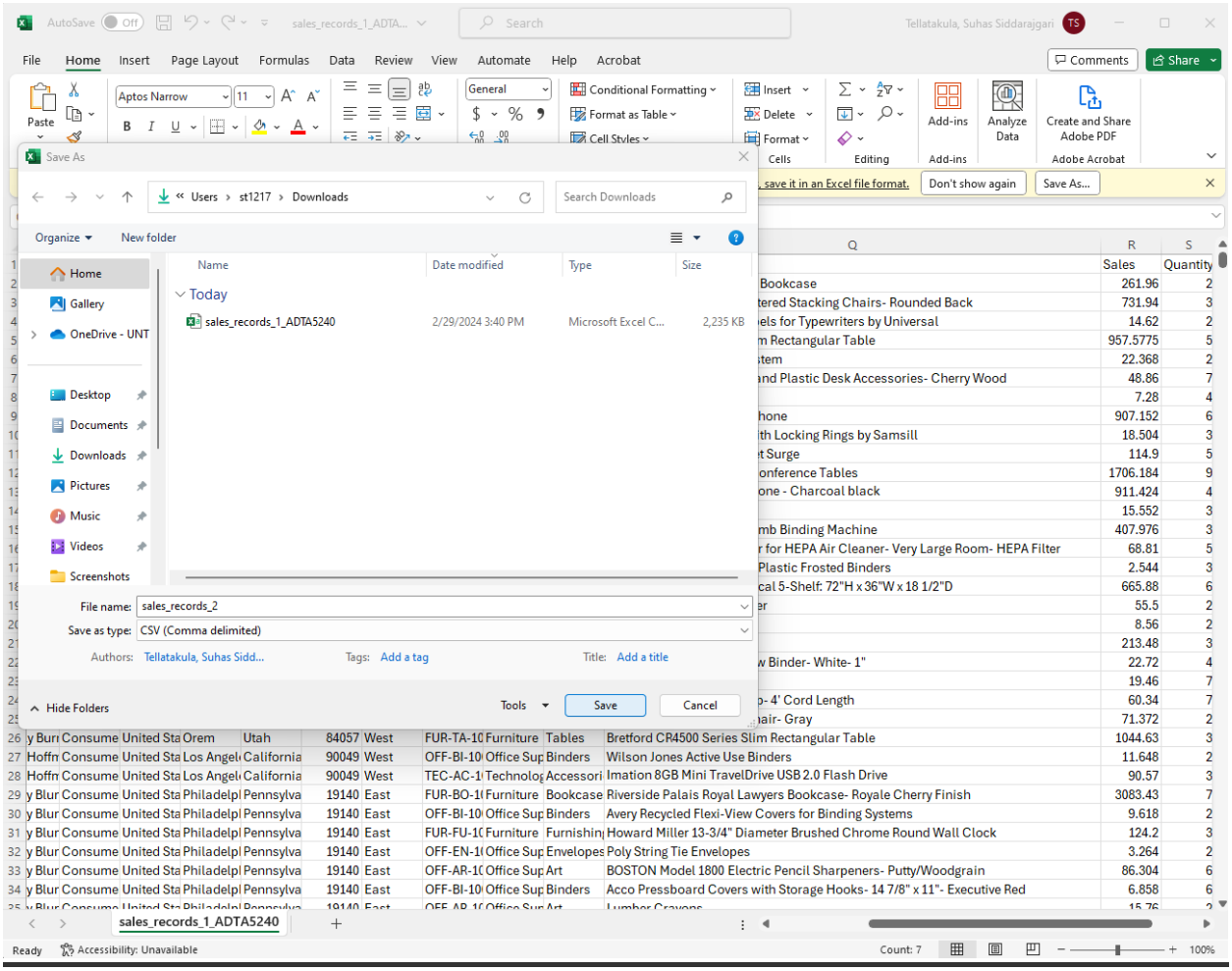
A small pop saying 3120 replacements were done is displayed as shown in below screenshot.



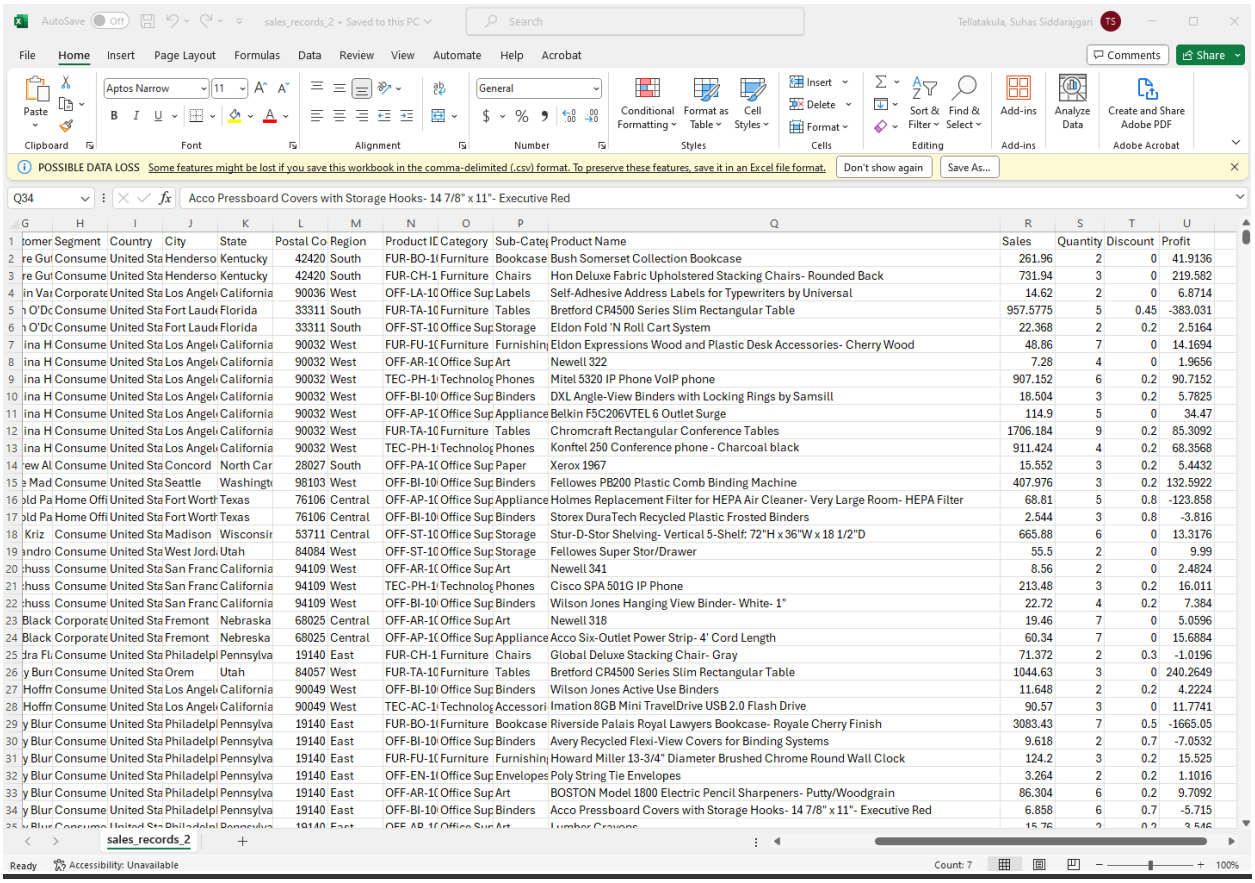
We can notice that all the commas are successfully replaced with hyphens in all highlighted cells under Product Name column as shown in below screenshot.



Next, in order to submit this cleaned dataset, I clicked on the save as option and saved it as sales_records_2.csv file as shown in below screenshot.



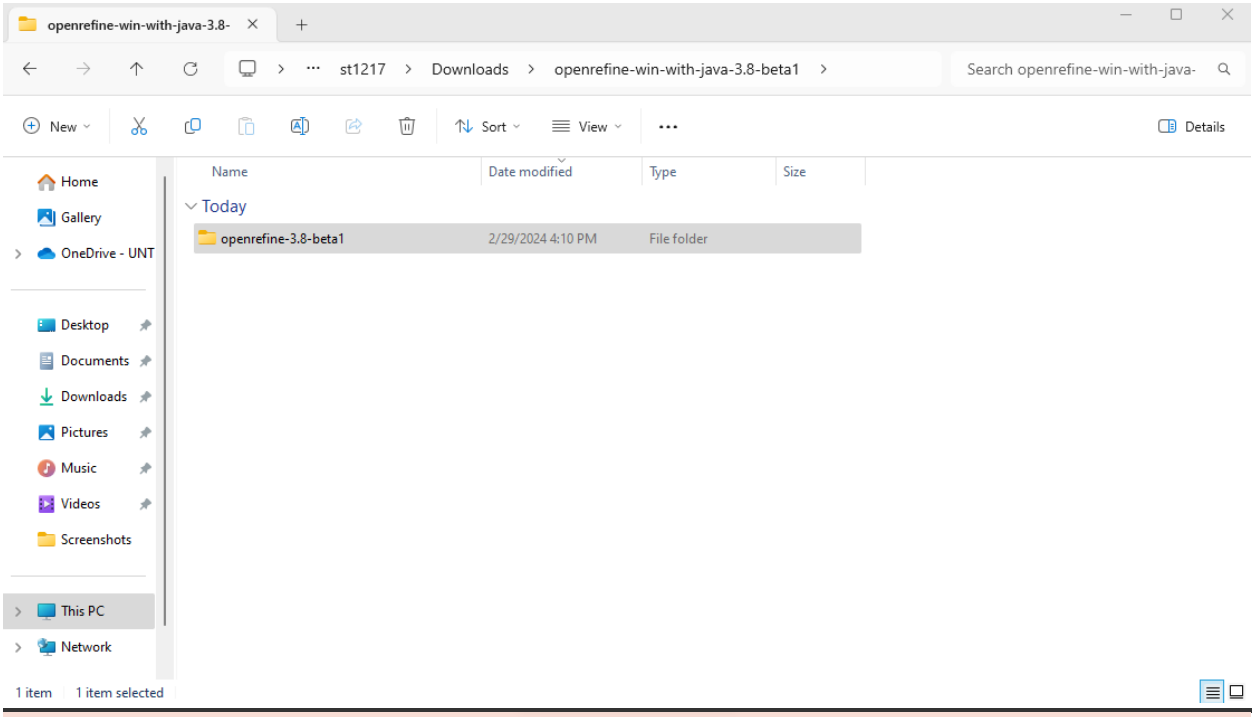
This screenshot shows the sales_records_2.csv file opened in Excel.

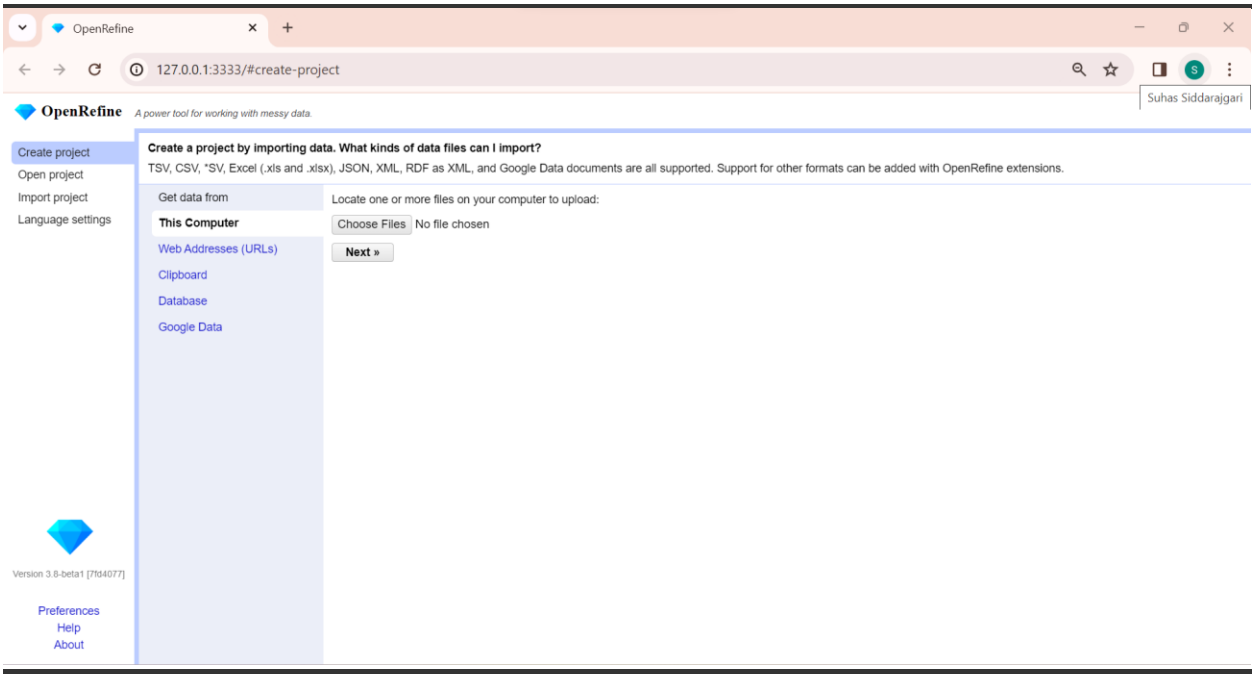


	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	tomer	Segment	Country	City	State	Postal Co	Region	Product ID	Category	Sub-Categ	Product Name	Sales	Quantity	Discount	Profit
2	re Gu	Consume	United Sta	Henderso	Kentucky	42420	South	FUR-BO-1(	Furniture	Bookcase	Bush Somerset Collection Bookcase	261.96	2	0	41.9136
3	re Gu	Consume	United Sta	Henderso	Kentucky	42420	South	FUR-CH-1	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs- Rounded Back	731.94	3	0	219.582
4	in Var	Corporate	United Sta	Los Angel	California	90036	West	OFF-LA-1(	Office Sup	Labels	Self-Adhesive Address Labels for Typewriters by Universal	14.62	2	0	6.8714
5	n O'Dc	Consume	United Sta	Fort Laud	Florida	33311	South	FUR-TA-1(	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	957.5775	5	0.45	-383.031
6	n O'Dc	Consume	United Sta	Fort Laud	Florida	33311	South	OFF-ST-1(	Office Sup	Storage	Eldon Fold 'N Roll Cart System	22.368	2	0.2	2.5164
7	ina H	Consume	United Sta	Los Angel	California	90032	West	FUR-FU-1(	Furniture	Furnishin	Eldon Expressions Wood and Plastic Desk Accessories- Cherry Wood	48.86	7	0	14.1694
8	ina H	Consume	United Sta	Los Angel	California	90032	West	OFF-AR-1(	Office Sup	Art	Newell 322	7.28	4	0	1.9656
9	ina H	Consume	United Sta	Los Angel	California	90032	West	TEC-PH-1(	Technolog	Phones	Mitel 5320 IP Phone VoIP phone	907.152	6	0.2	90.7152
10	ina H	Consume	United Sta	Los Angel	California	90032	West	OFF-BI-10	Office Sup	Binders	DXL Angle-View Binders with Locking Rings by Samsill	18.504	3	0.2	5.7825
11	ina H	Consume	United Sta	Los Angel	California	90032	West	OFF-AP-1(	Office Sup	Appliance	Belkin F5C206VTEL 6 Outlet Surge	114.9	5	0	34.47
12	ina H	Consume	United Sta	Los Angel	California	90032	West	FUR-TA-1(	Furniture	Tables	Chromcraft Rectangular Conference Tables	1706.184	9	0.2	85.3092
13	ina H	Consume	United Sta	Los Angel	California	90032	West	TEC-PH-1(	Technolog	Phones	Konftel 250 Conference phone - Charcoal black	911.424	4	0.2	68.3568
14	ew Al	Consume	United Sta	Concord	North Car	28027	South	OFF-PA-1(	Office Sup	Paper	Xerox 1967	15.552	3	0.2	5.4432
15	e Mad	Consume	United Sta	Seattle	Washingt	98103	West	OFF-BI-10	Office Sup	Binders	Fellowes PB200 Plastic Comb Binding Machine	407.976	3	0.2	132.5922
16	old Pa	Home Offi	United Sta	Fort Worth	Texas	76106	Central	OFF-AP-1(	Office Sup	Appliance	Holmes Replacement Filter for HEPA Air Cleaner- Very Large Room- HEPA Filter	68.81	5	0.8	-123.858
17	old Pa	Home Offi	United Sta	Fort Worth	Texas	76106	Central	OFF-BI-10	Office Sup	Binders	Storex DuraTech Recycled Plastic Frosted Binders	2.544	3	0.8	-3.816
18	Kriz	Consume	United Sta	Madison	Wisconsir	53711	Central	OFF-ST-1(	Office Sup	Storage	Stur-D-Stor Shelving- Vertical 5-Shelf: 72"H x 36"W x 18 1/2"D	665.88	6	0	13.3176
19	andro	Consume	United Sta	West Jord	Utah	84084	West	OFF-ST-1(	Office Sup	Storage	Fellowes Super Stor/Drawer	55.5	2	0	9.99
20	huss	Consume	United Sta	San Franc	California	94109	West	OFF-AR-1(	Office Sup	Art	Newell 341	8.56	2	0	2.4824
21	huss	Consume	United Sta	San Franc	California	94109	West	TEC-PH-1(	Technolog	Phones	Cisco SPA 501G IP Phone	213.48	3	0.2	16.011
22	huss	Consume	United Sta	San Franc	California	94109	West	OFF-BI-10	Office Sup	Binders	Wilson Jones Hanging View Binder- White- 1"	22.72	4	0.2	7.384
23	Black	Corporate	United Sta	Fremont	Nebraska	68025	Central	OFF-AR-1(	Office Sup	Art	Newell 318	19.46	7	0	5.0596
24	Black	Corporate	United Sta	Fremont	Nebreska	68025	Central	OFF-AP-1(	Office Sup	Appliance	Acco Six-Outlet Power Strip- 4' Cord Length	60.34	7	0	15.6884
25	dra Fl	Consume	United Sta	Philadelp	Pennsylva	19140	East	FUR-CH-1	Furniture	Chairs	Global Deluxe Stacking Chair- Gray	71.372	2	0.3	-1.0196
26	y Burt	Consume	United Sta	Orem	Utah	84057	West	FUR-TA-1(	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	1044.63	3	0	240.2649
27	Hoffm	Consume	United Sta	Los Angel	California	90049	West	OFF-BI-10	Office Sup	Binders	Wilson Jones Active Use Binders	11.648	2	0.2	4.2224
28	Hoffm	Consume	United Sta	Los Angel	California	90049	West	TEC-AC-1(	Technolog	Accessori	Imation 8GB Mini TravelDrive USB 2.0 Flash Drive	90.57	3	0	11.7741
29	y Blur	Consume	United Sta	Philadelp	Pennsylva	19140	East	FUR-BO-1(	Furniture	Bookcase	Riverside Palais Royal Lawyers Bookcase- Royale Cherry Finish	3083.43	7	0.5	-1665.05
30	y Blur	Consume	United Sta	Philadelp	Pennsylva	19140	East	OFF-BI-10	Office Sup	Binders	Avery Recycled Flexi-View Covers for Binding Systems	9.618	2	0.7	-7.0532
31	y Blur	Consume	United Sta	Philadelp	Pennsylva	19140	East	FUR-FU-1(	Furniture	Furnishin	Howard Miller 13-3/4" Diameter Brushed Chrome Round Wall Clock	124.2	3	0.2	15.525
32	y Blur	Consume	United Sta	Philadelp	Pennsylva	19140	East	OFF-EN-1(	Office Sup	Envelopes	Poly String Tie Envelopes	3.264	2	0.2	1.1016
33	y Blur	Consume	United Sta	Philadelp	Pennsylva	19140	East	OFF-AR-1(	Office Sup	Art	BOSTON Model 1800 Electric Pencil Sharpeners- Putty/Woodgrain	86.304	6	0.2	9.7092
34	y Blur	Consume	United Sta	Philadelp	Pennsylva	19140	East	OFF-BI-10	Office Sup	Binders	Acco Pressboard Covers with Storage Hooks- 14 7/8" x 11"- Executive Red	6.858	6	0.7	-5.715

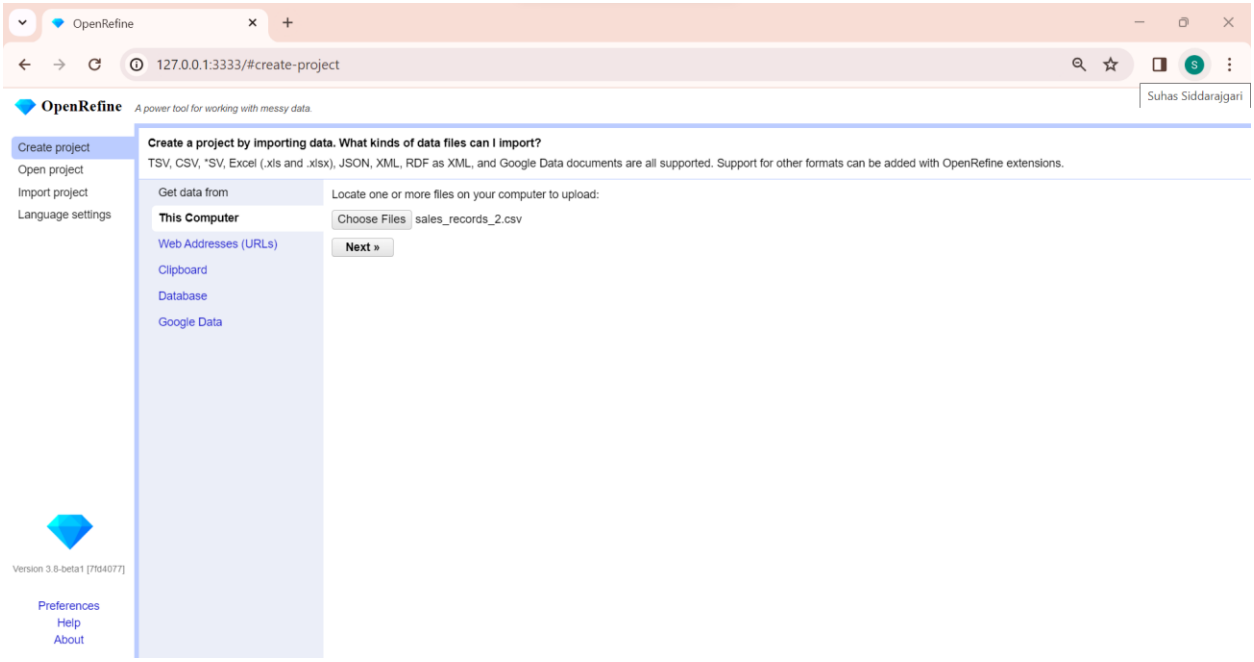
Part 3a – OpenRefine Data Preprocessing

For this part of data preprocessing, I will use OpenRefine. First, I downloaded and installed the Openrefine tool in my local and opened it as shown in below 2 screenshots.

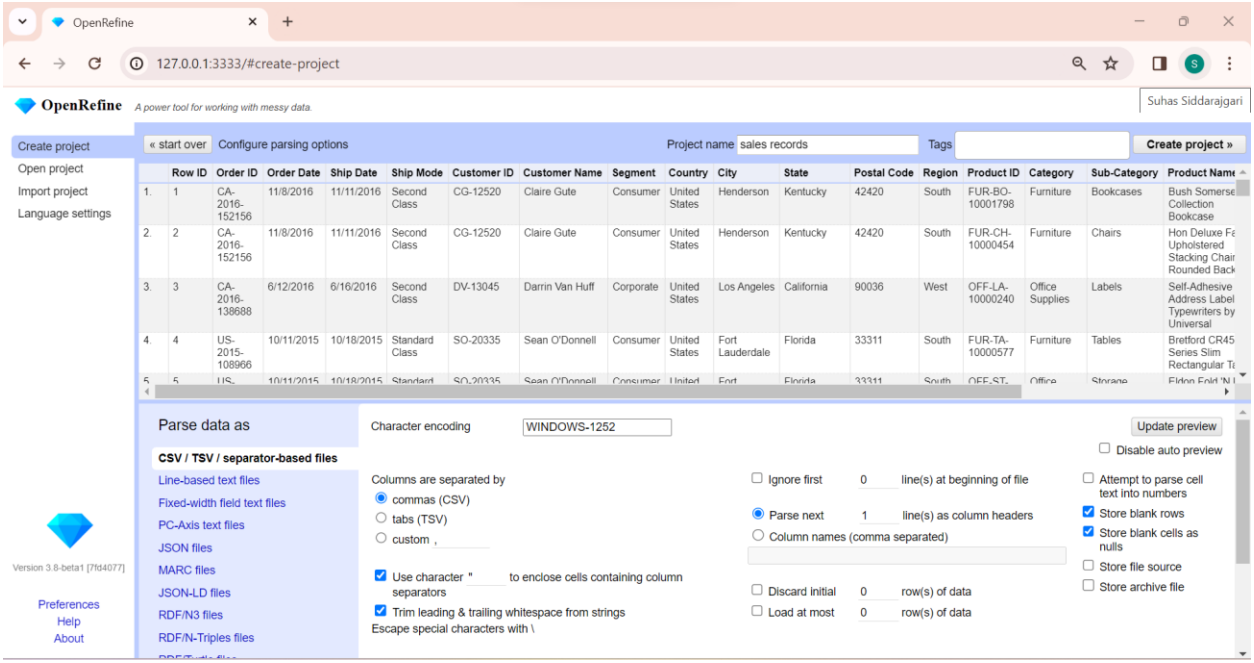




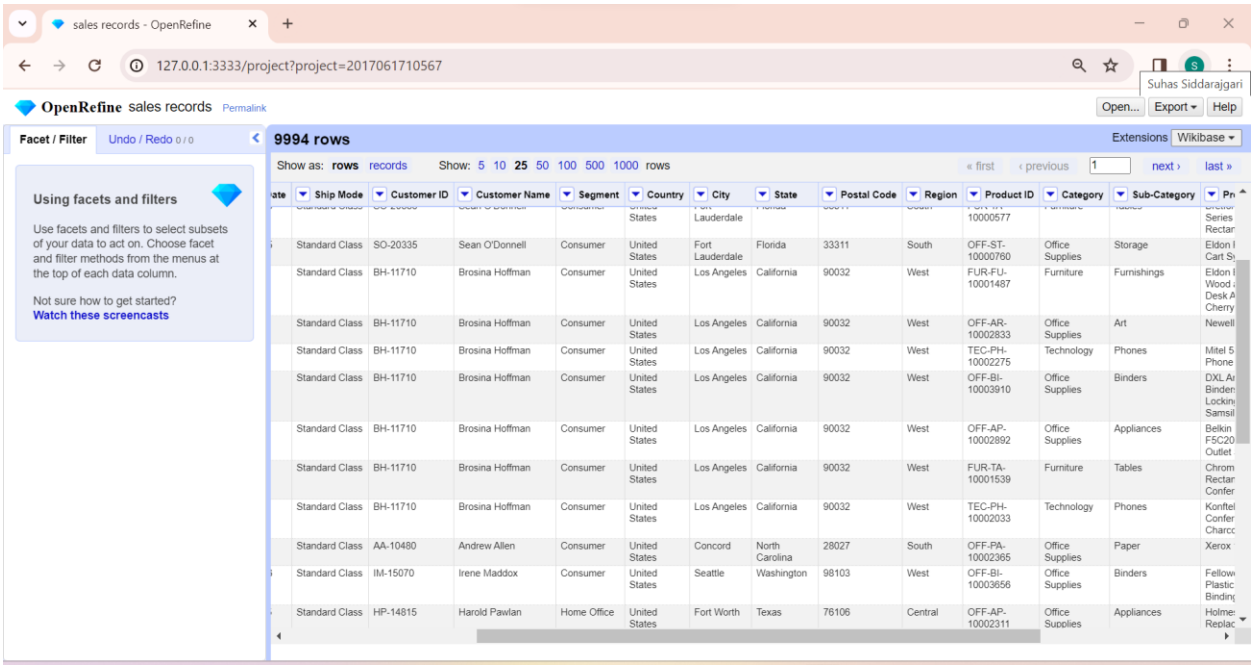
Next, I uploaded the sales_records_2.csv file into OpenRefine as shown in below screenshot and clicked on Next.



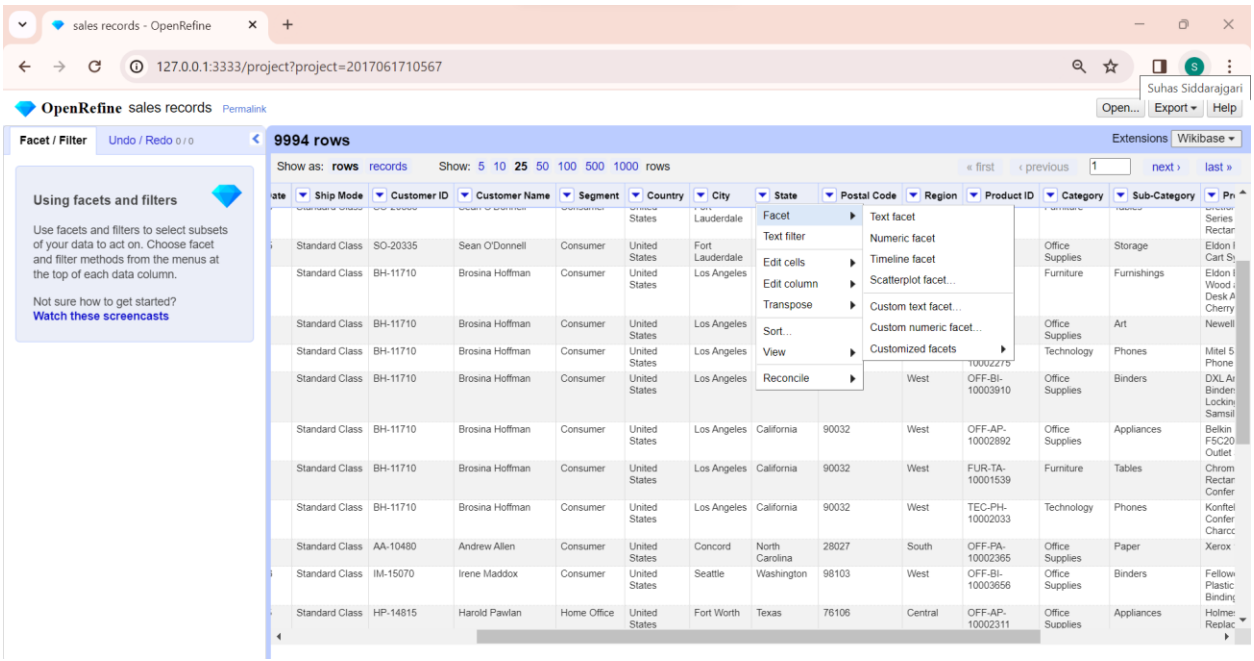
Here, I gave the project name as sales records and left the default properties as shown in below screenshot and clicked on the create project button.



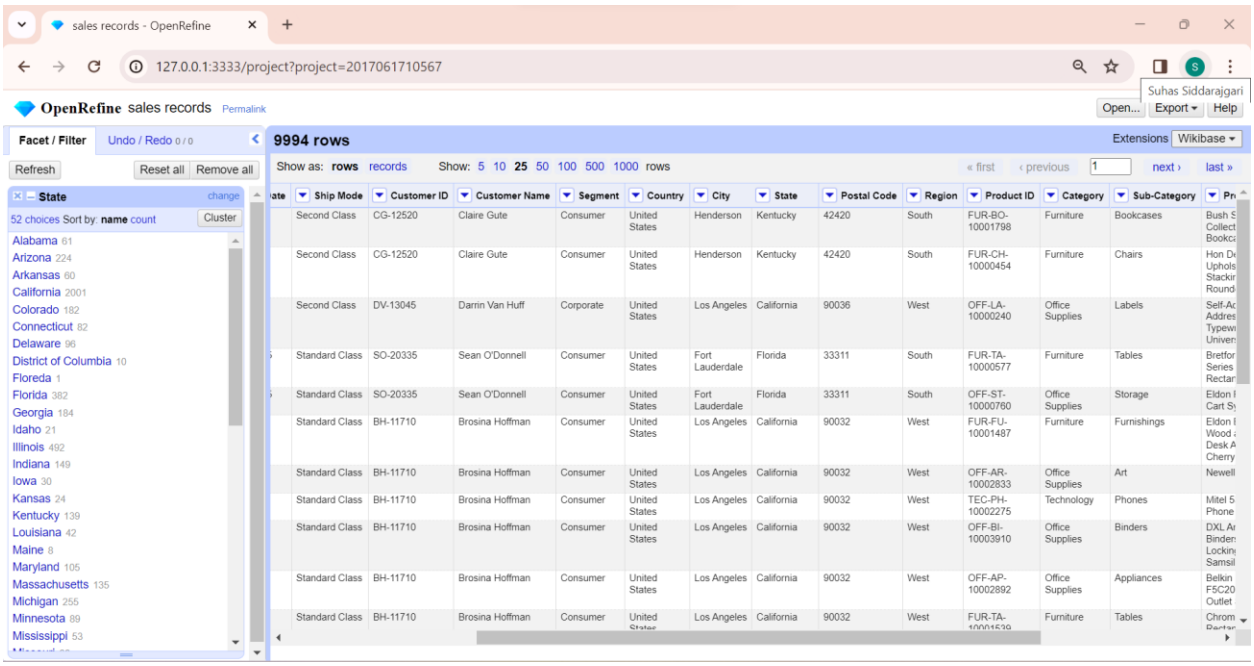
Once the project is created successfully, we can see the data has 9994 records of sales as shown in below screenshot.



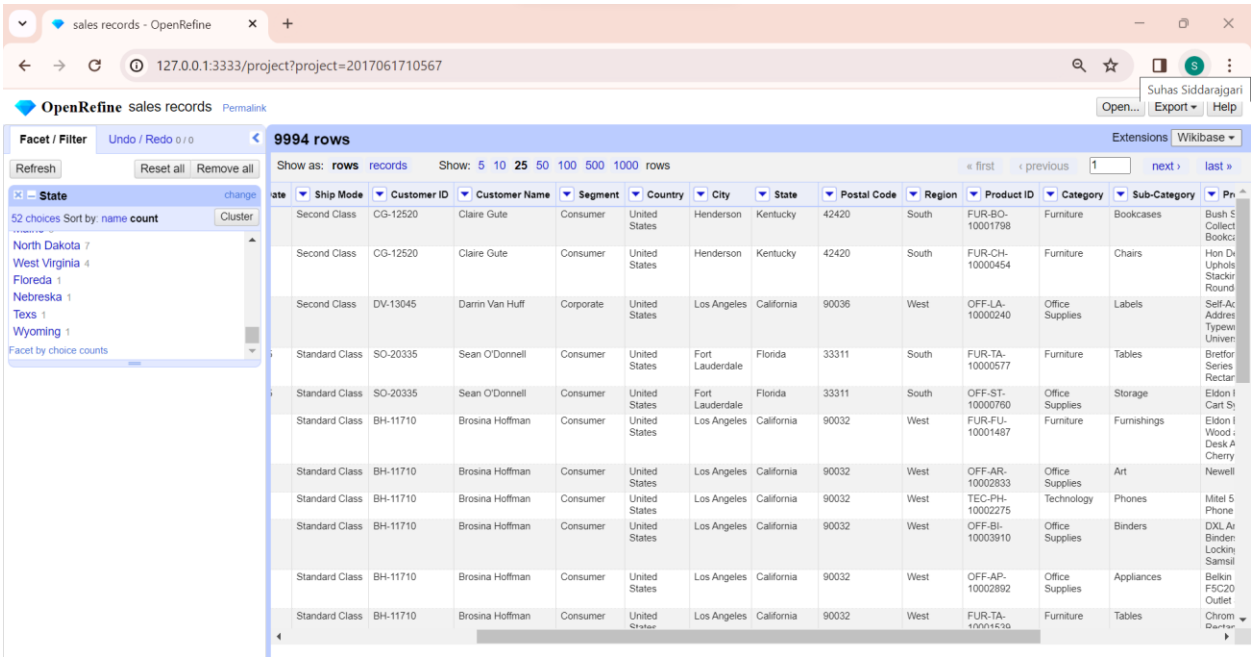
In order to clean the values in State column, I clicked on the menu option for State and chose facet -> Text Facet as shown in below screenshot.



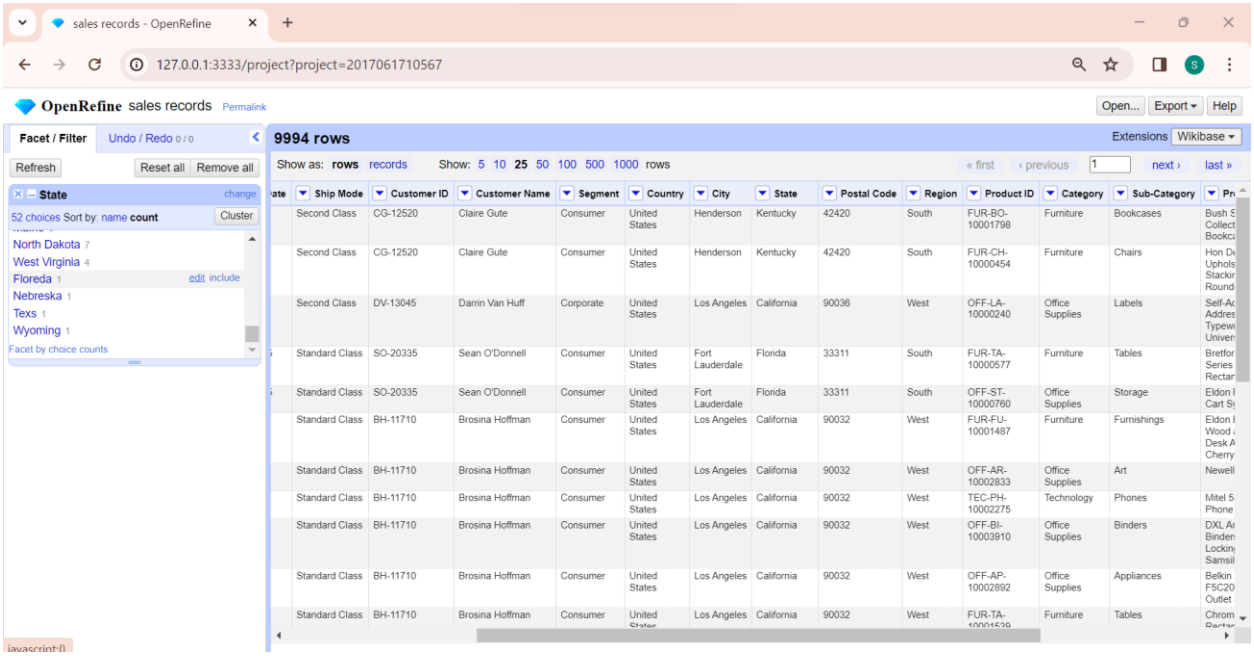
We can see the text facet results in the left tab with frequencies sorted by names as shown in below screenshot.



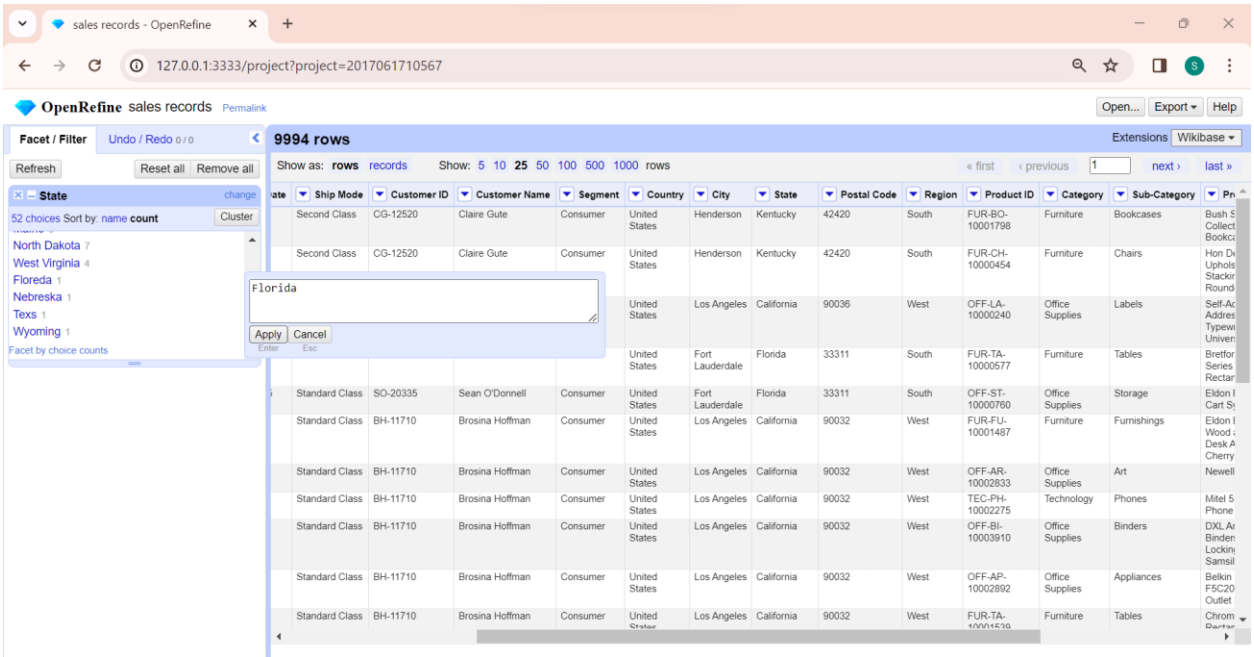
As there are few typo occurrences, I clicked on the sort by count which shows the typo occurrences like “Floreda, Nebreska, Texs” as shown in the below screenshot.



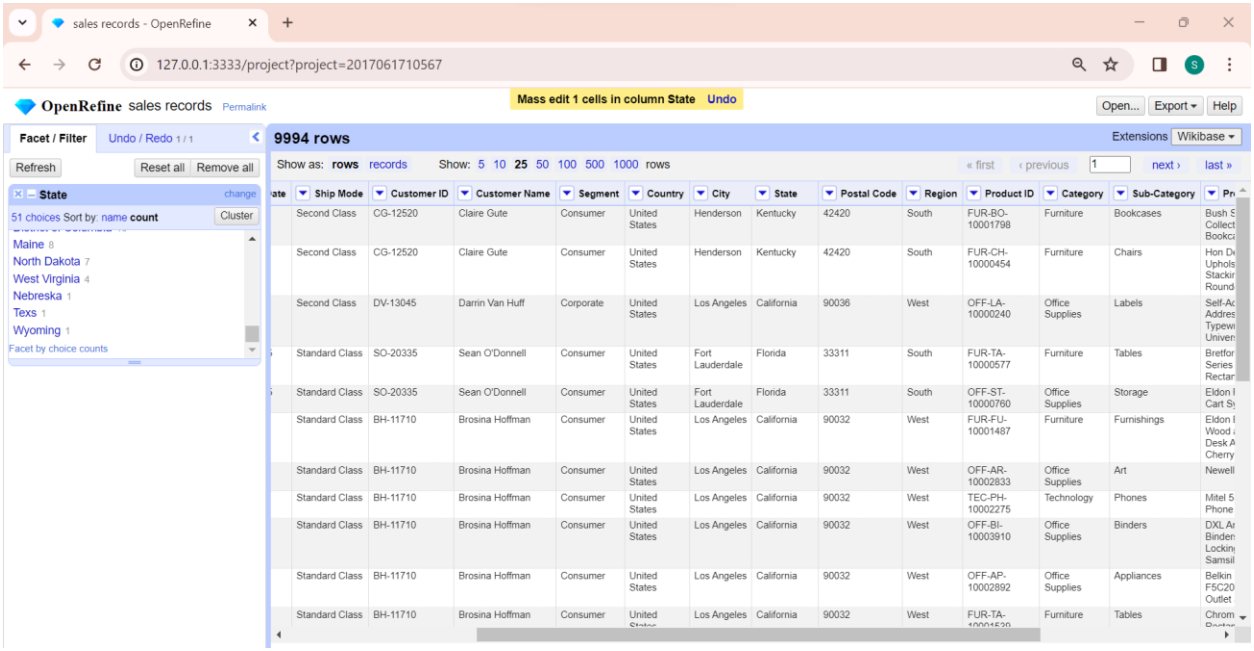
Since there are only 3 typos, I used the edit option in front of these typos as shown in below screenshot.



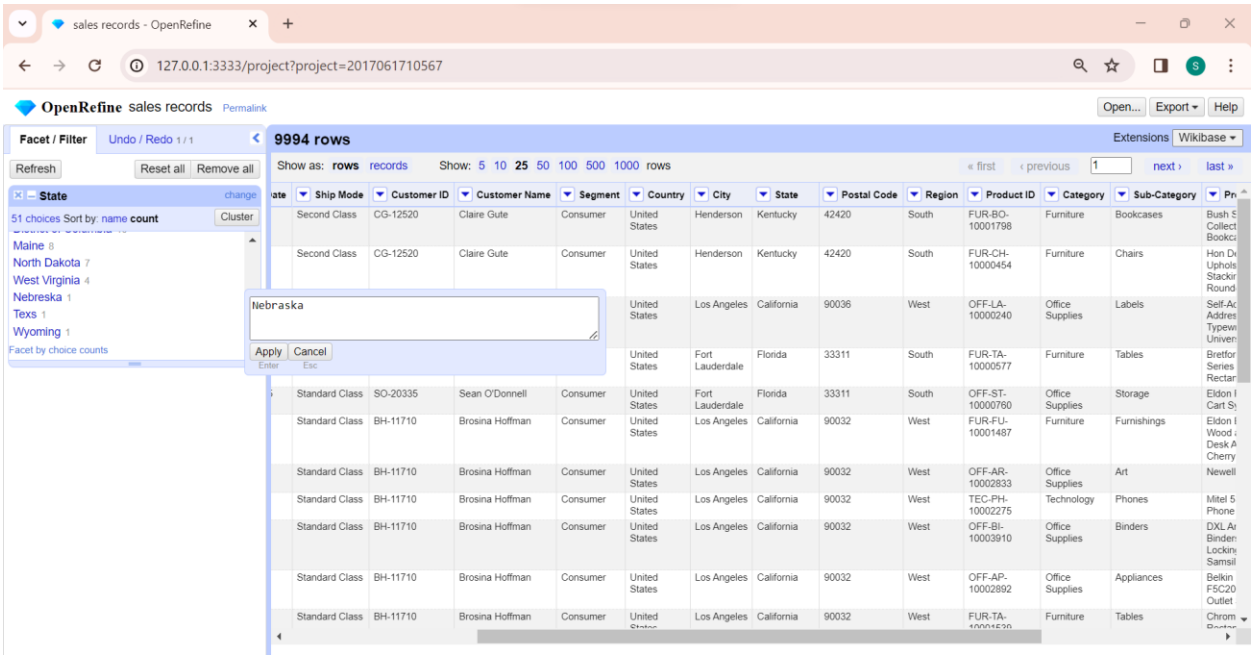
Then, I corrected the spelling and clicked on the apply typos as shown in below screenshot.



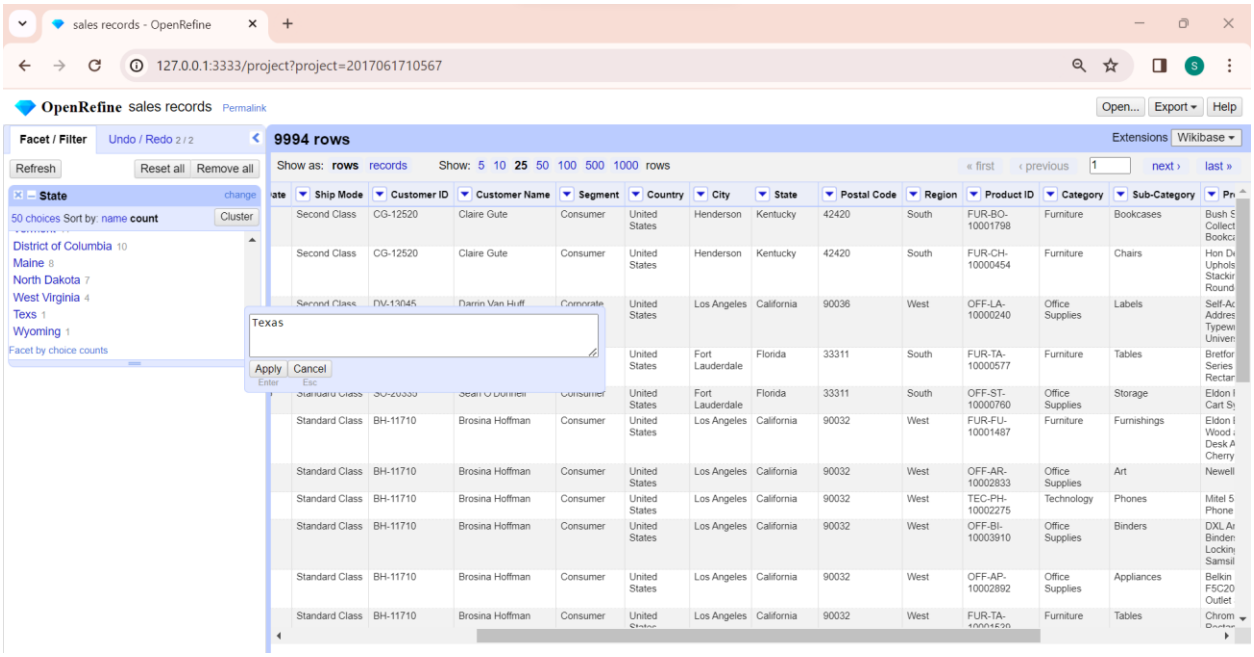
We can see that the typo has been fixed and is not showing under facet results anymore as shown in below screenshot.



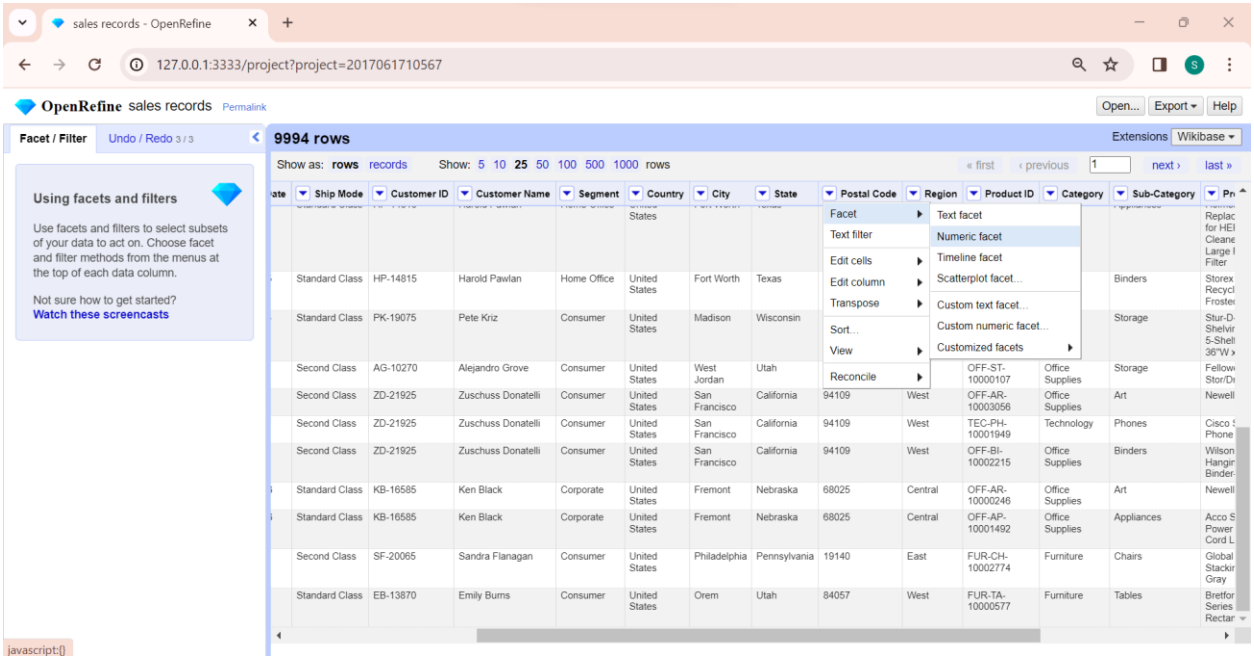
Similarly, I fixed the spelling of Nebreska as shown in the screenshot below.



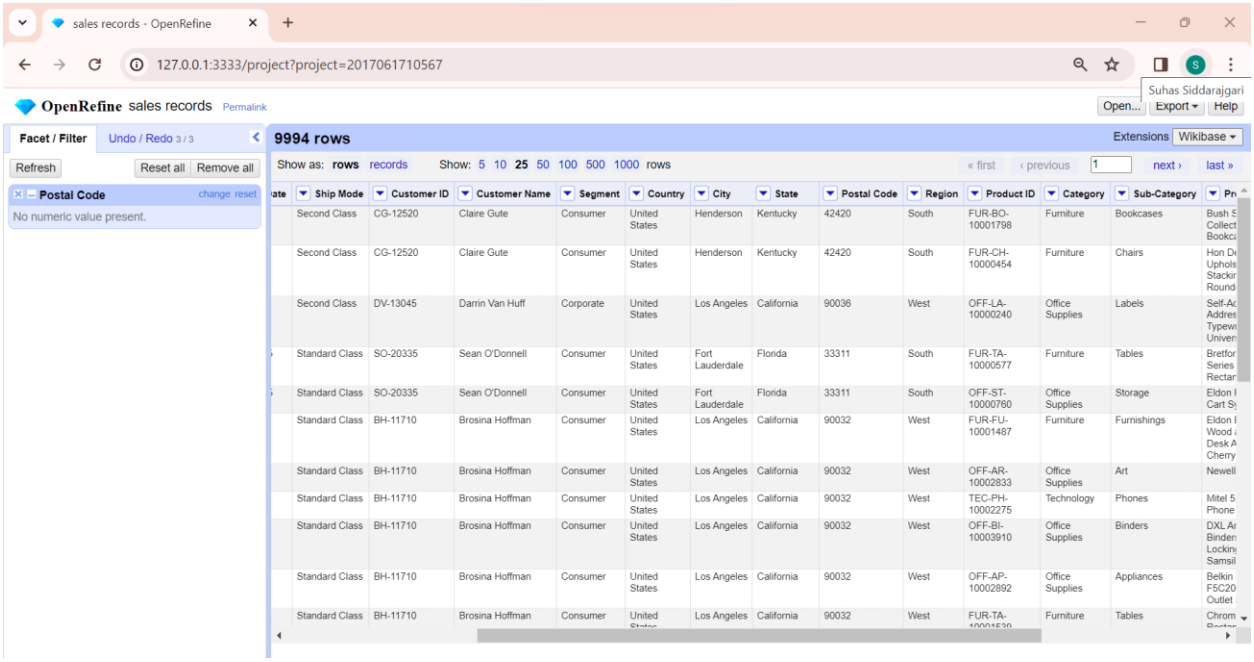
Similarly, I fixed the spelling of Texs as shown in the screenshot below.



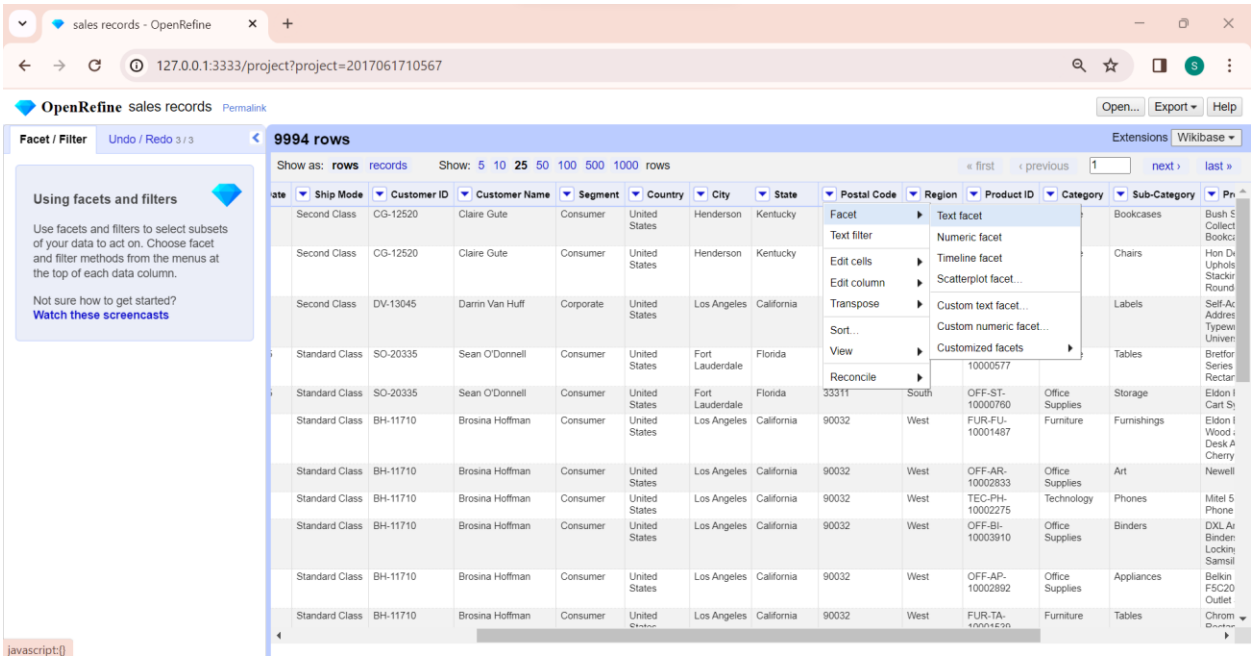
Next, I clicked on the menu of Postal code and clicked on the facet-> Numeric facet as shown in the screenshot below.



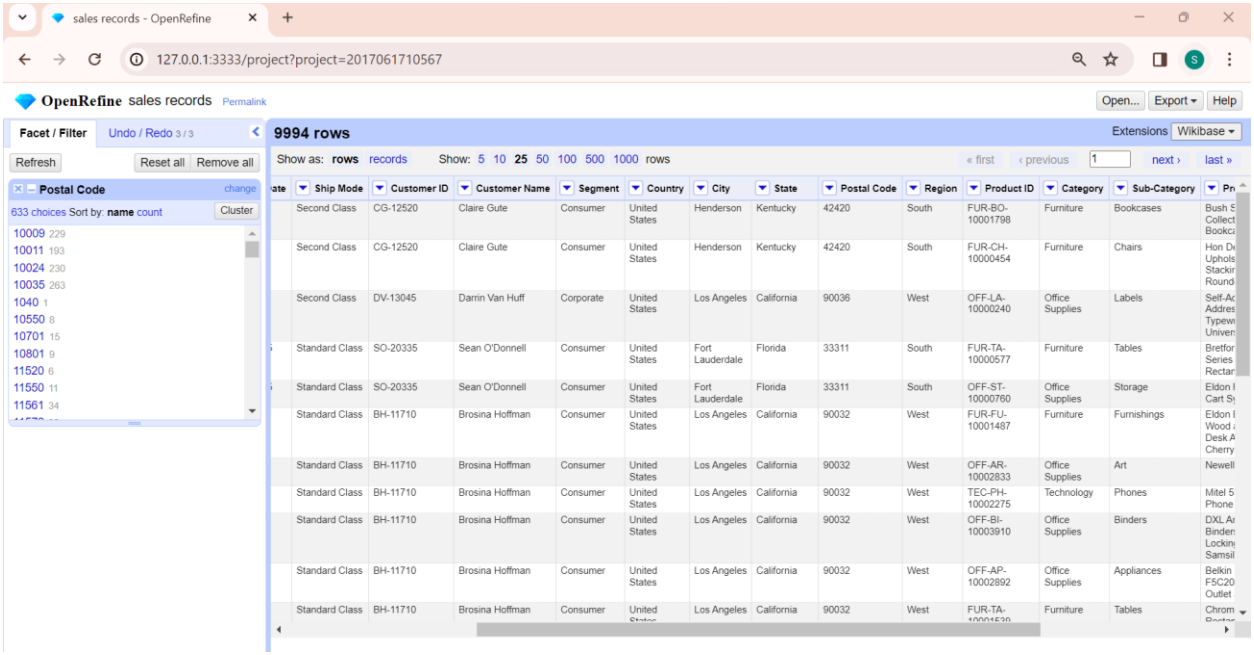
We can see that there are no numeric values in postal code as shown in the screenshot below which is a mistake as zip codes should be numeric data.



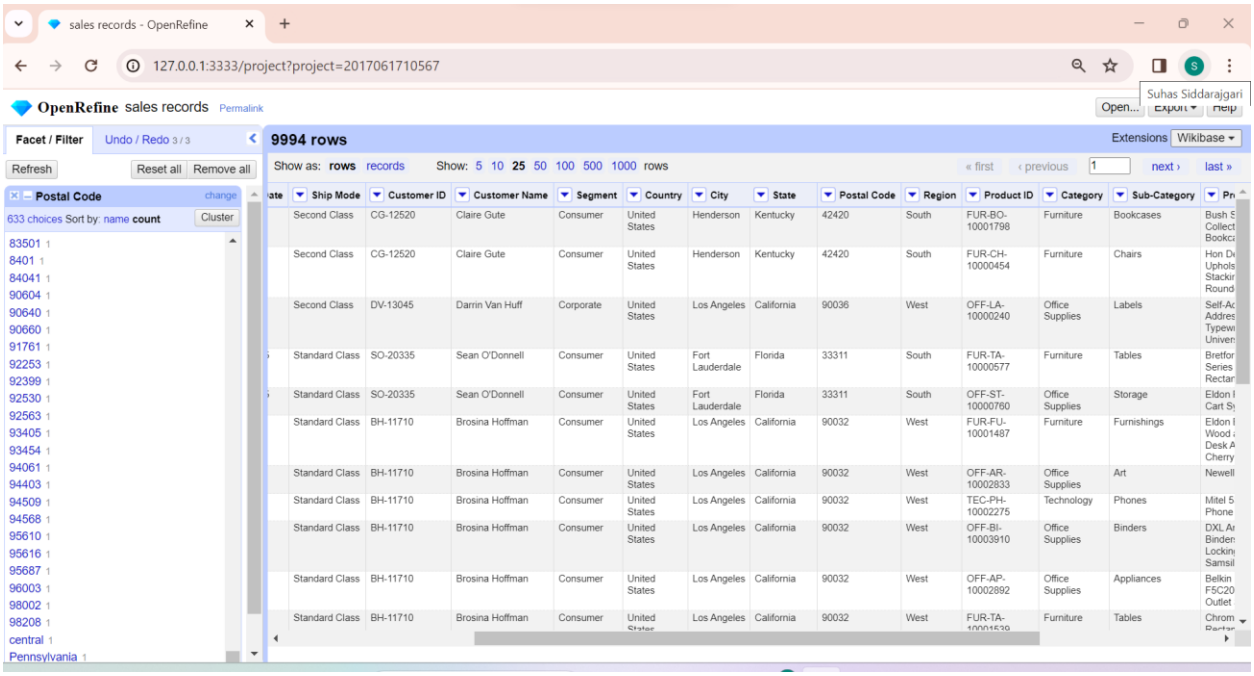
Next, I clicked on the menu of Postal code and clicked on the facet-> Text facet as shown in the screenshot below.



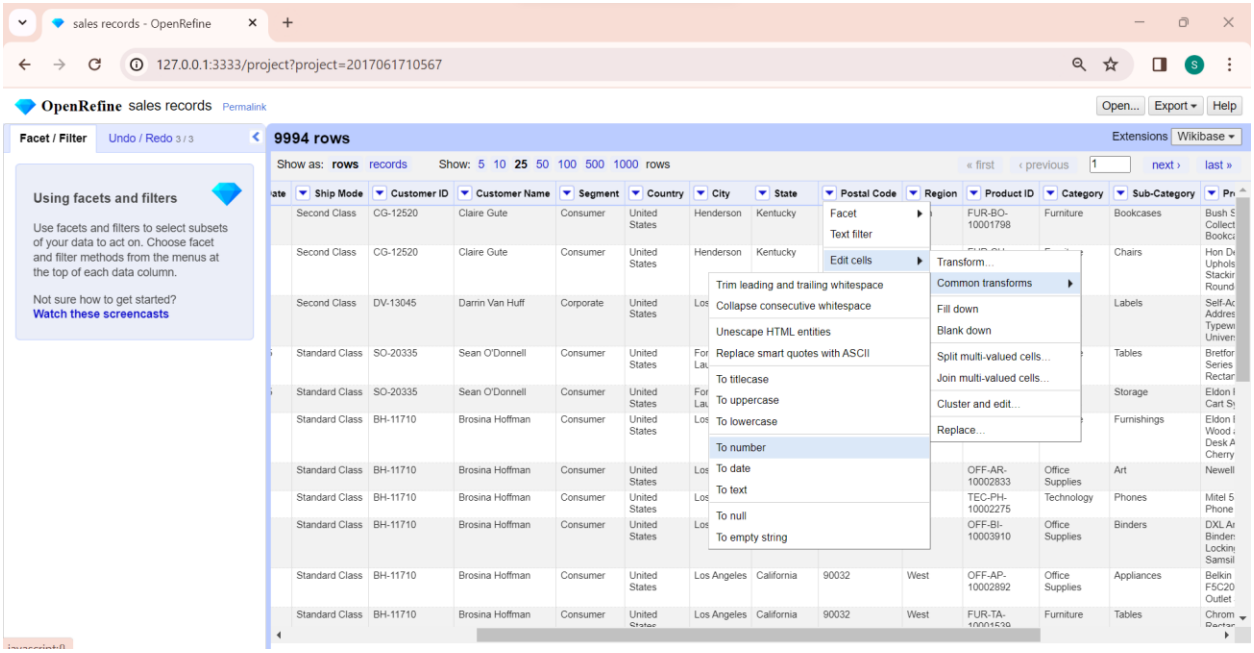
We can see the results of the text facets of the zip codes on the left side of the screenshot below.



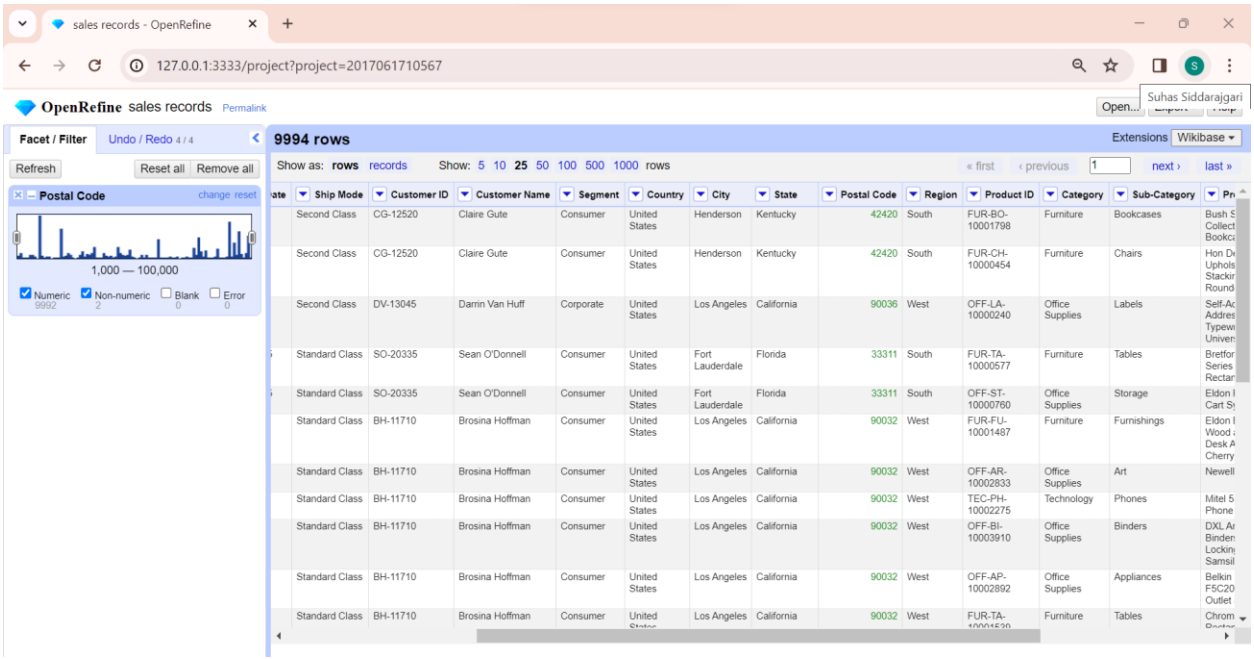
We can notice that there are 2 text strings entered under postal codes – central and Pennsylvania which need to be fixed as shown in the below screenshot.



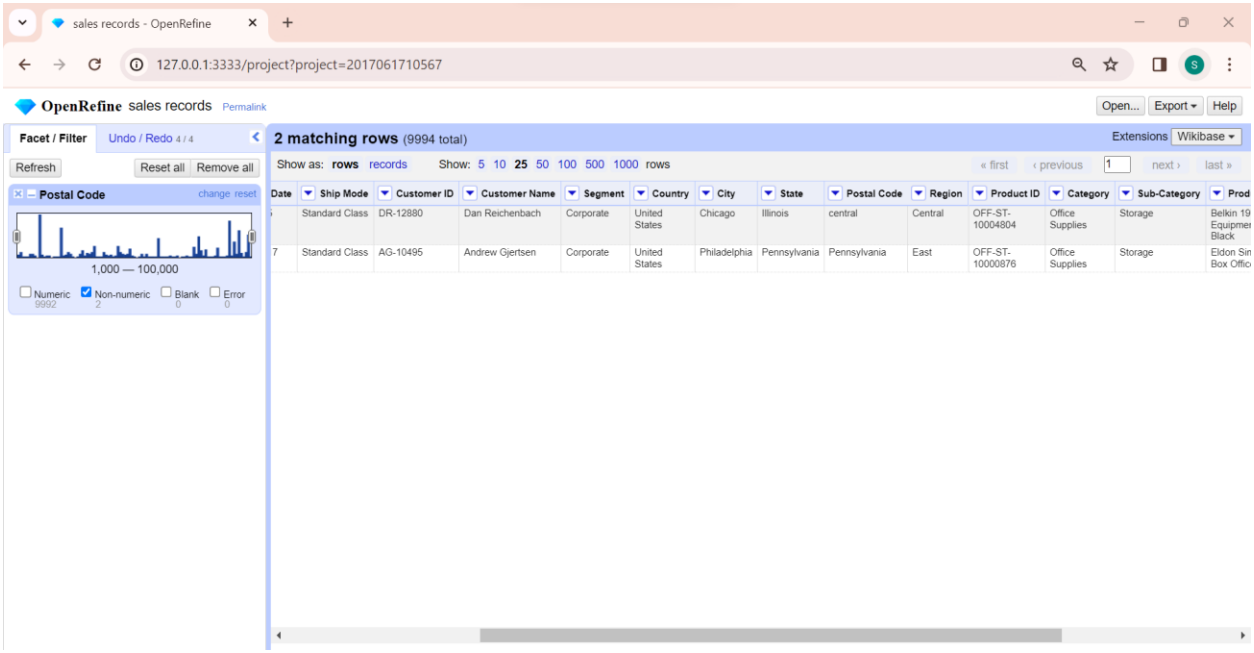
In order to convert these text data into numeric data, I clicked on the menu-> Edit cells-> common transforms-> To number as shown in the below screenshot.



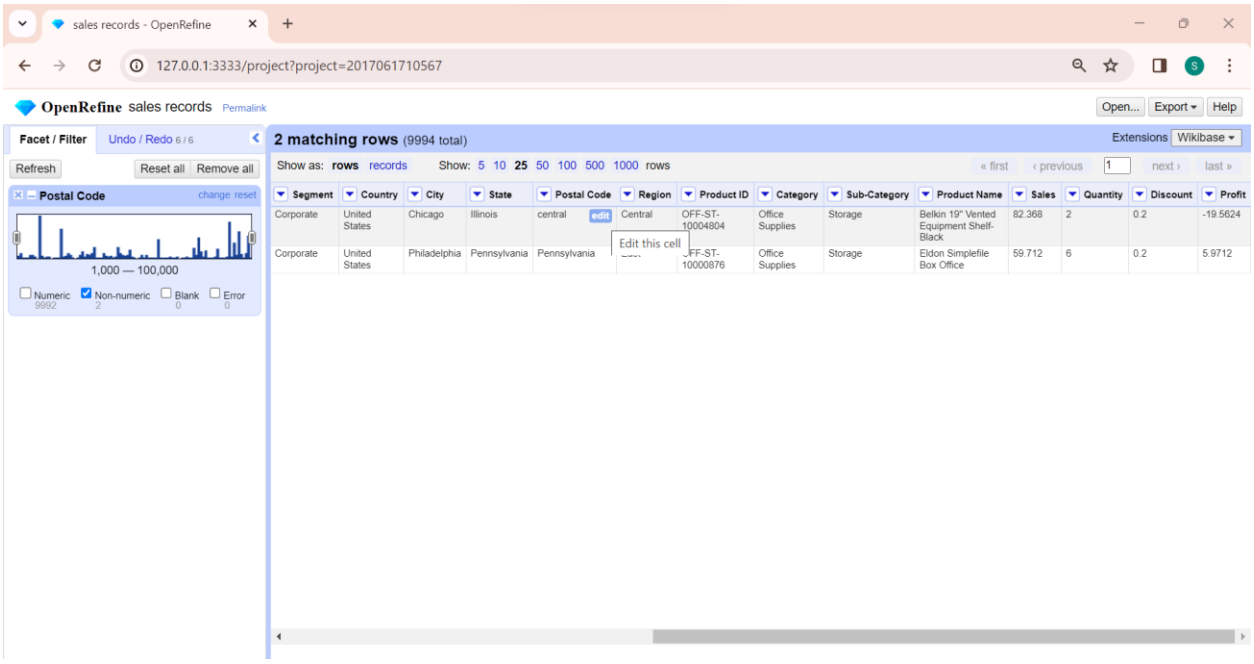
We can see that the text data are successfully converted to numeric values as I checked the numeric facets again as shown in the screenshot below.



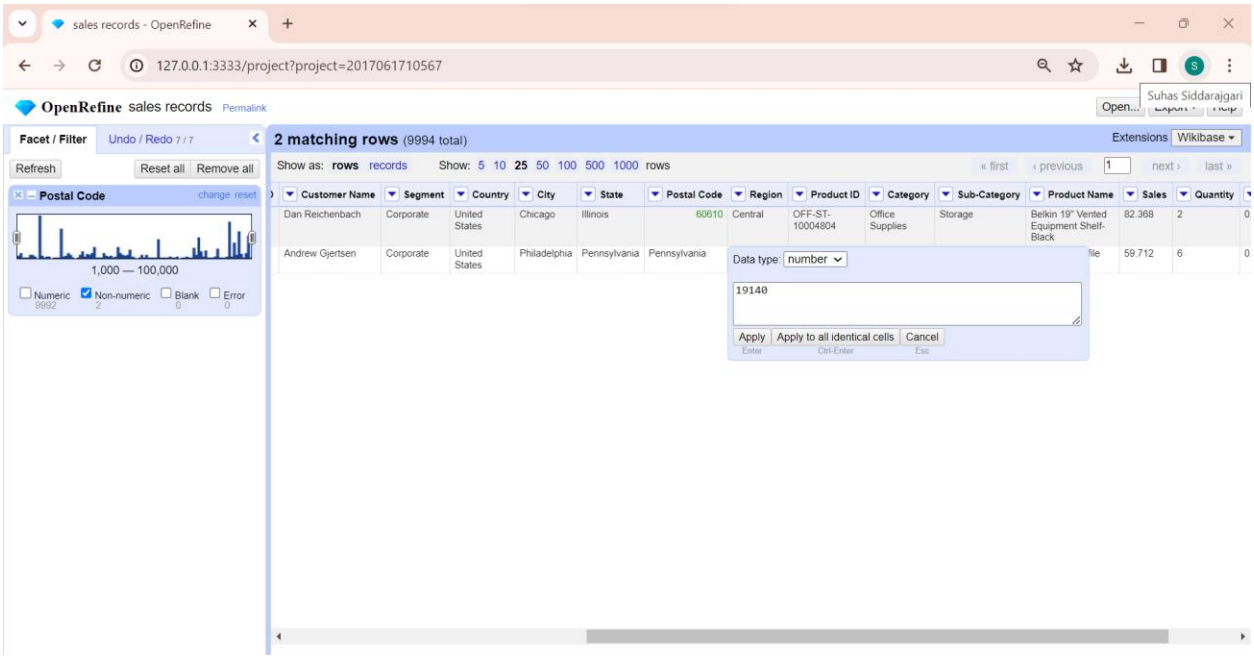
In order to fix the 2 nonnumeric data, I unchecked the numeric so that I can edit those values as shown in the screenshot below.



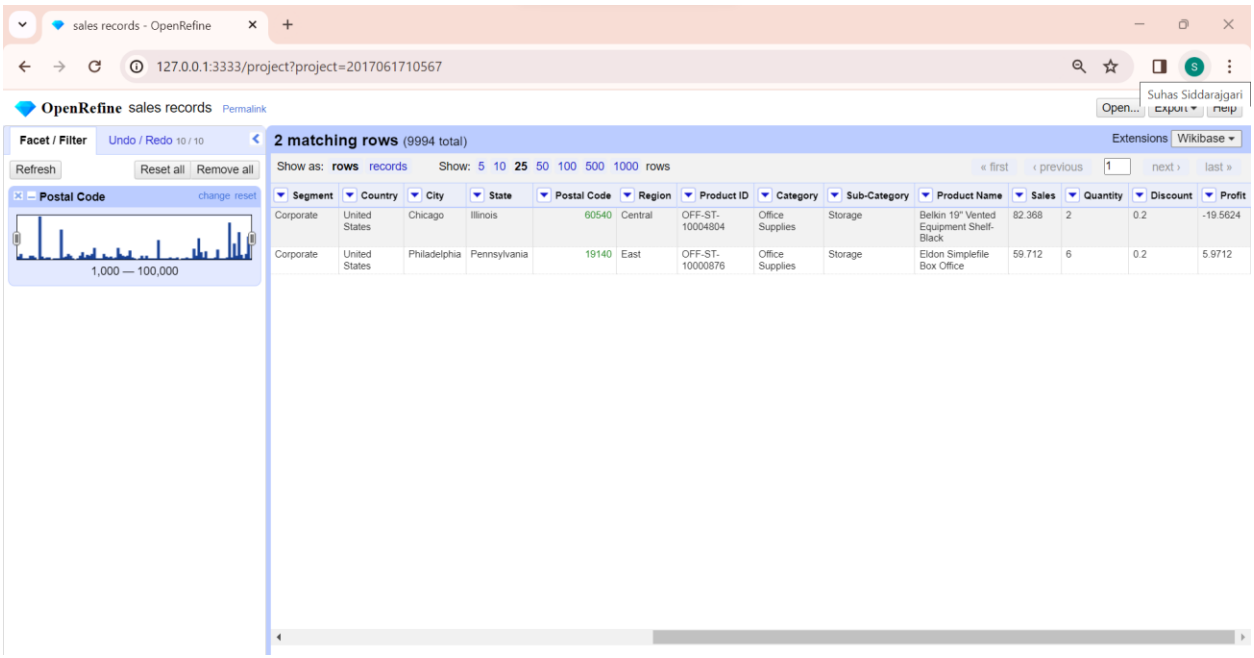
Here, I will use the edit option in the cell to fix those values which is shown in the screenshot below.



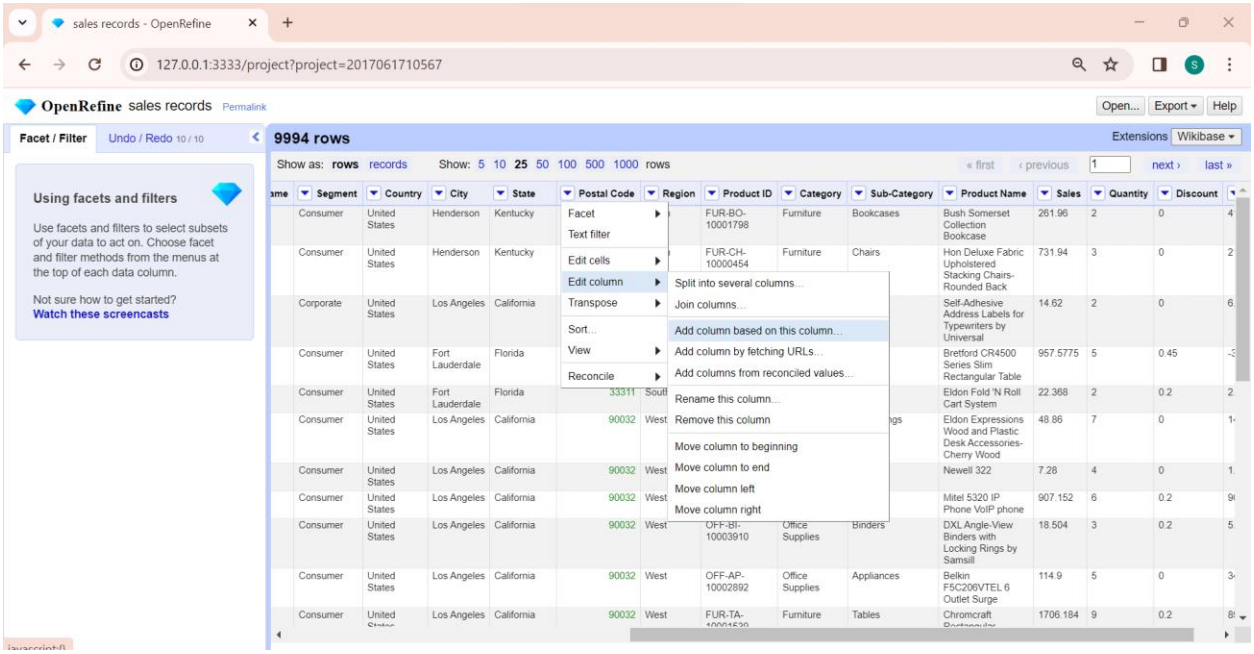
I referred the data and looked for the zip code of the city Chicago in Illinois state which is 60610 and 19140 for Philadelphia city in Pennsylvania state and chose the data type as number as shown in the screenshot below.



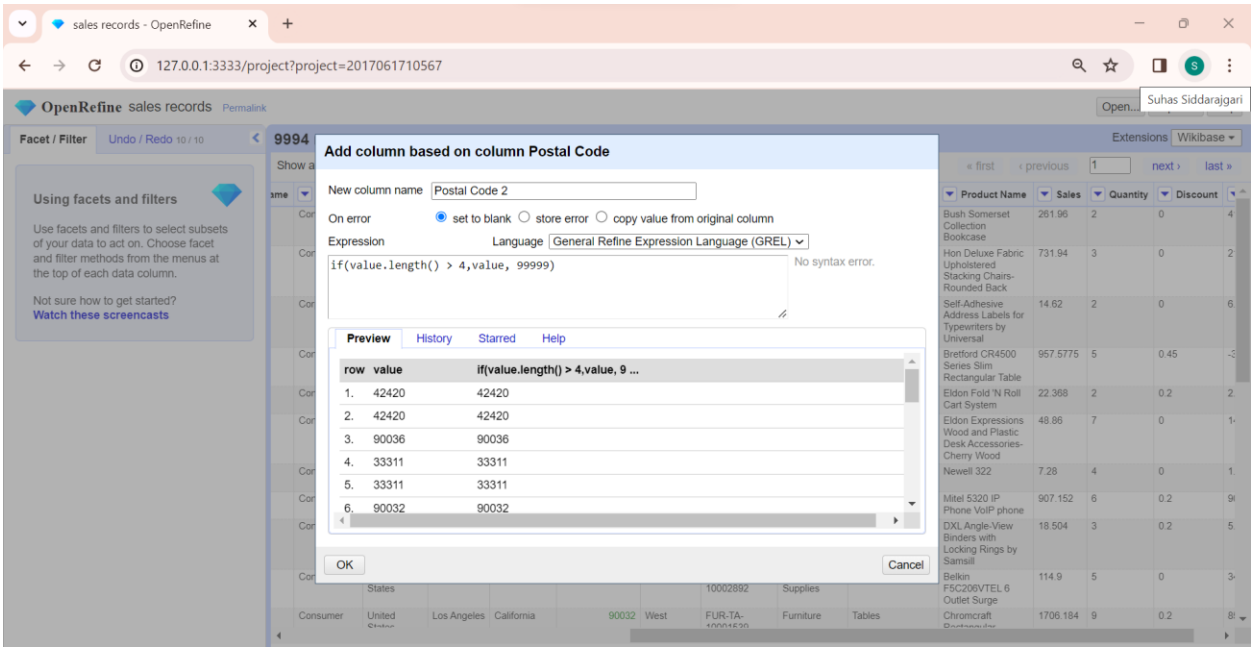
We can see that the 2 nonnumeric values are now fixed as shown in below screenshot.



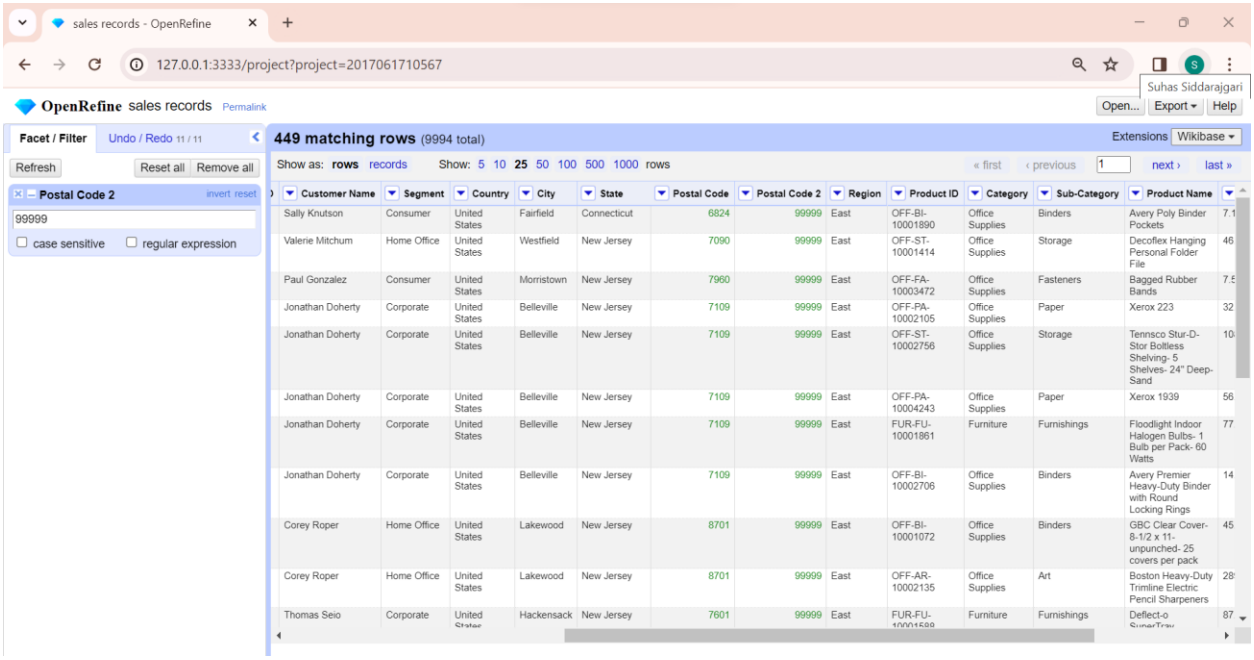
Next, I clicked on the menu of Postal code-> Edit column-> Add column based on this column as shown in the below screenshot.



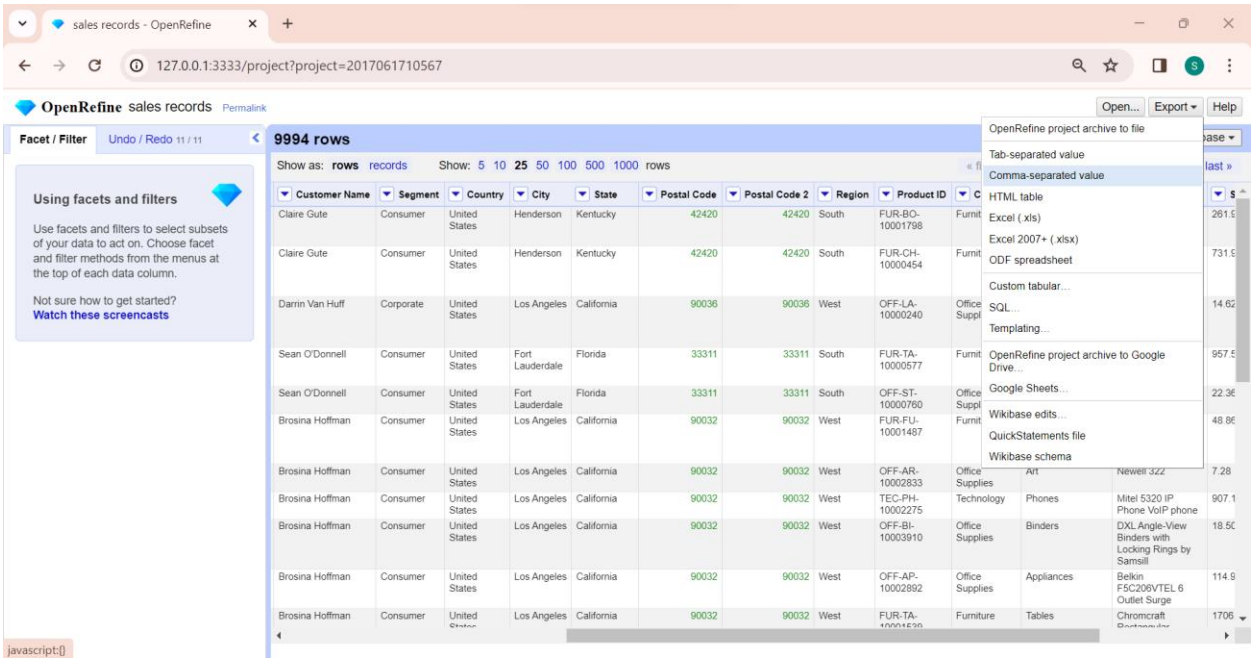
In the add column form, I gave new column name as Postal code 2 and entered expression as “if(value.length() > 4, value, 99999) as there are few records which are not 5 digits so that if they are invalid, this column can serve as correct zip code as shown in below screenshot.



In this screenshot, we can see that for records with postal codes other than 5 digits, Postal Code 2 column has value of 99999 leaving the rest as it is.



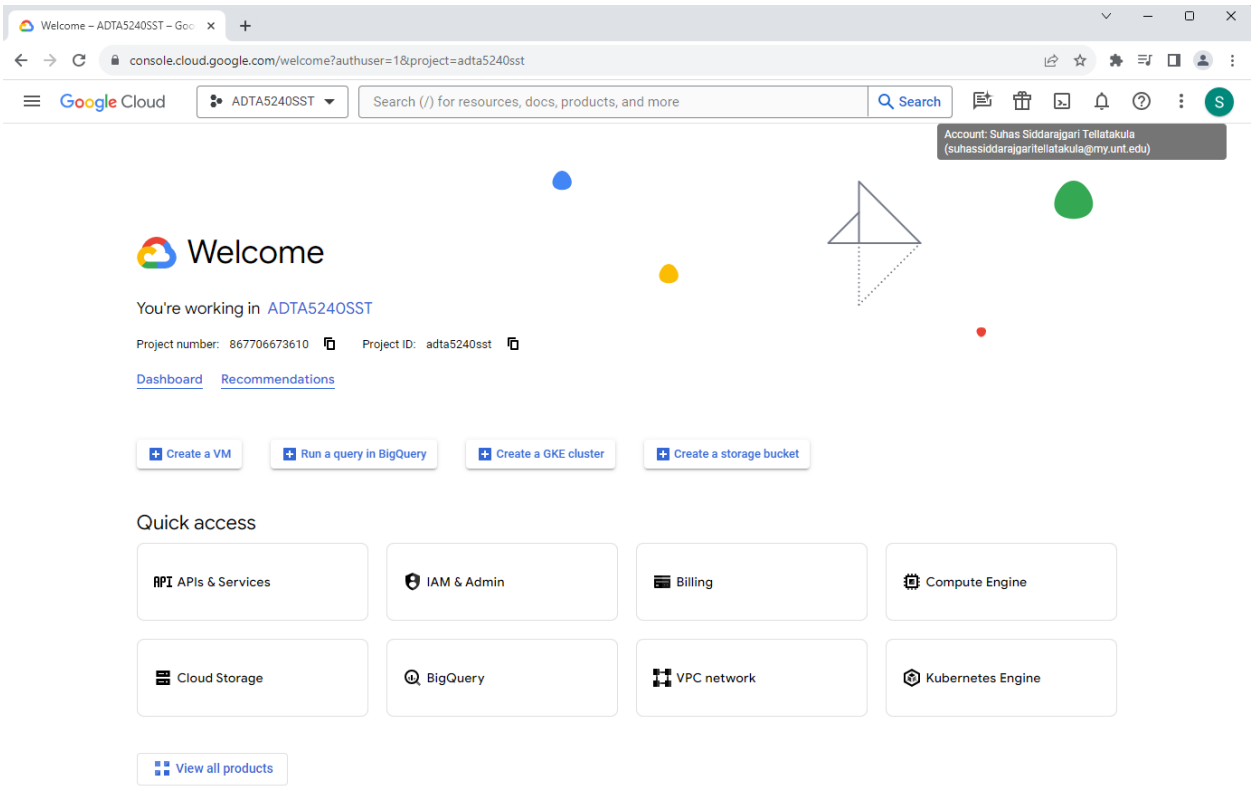
Next, I clicked on the export button on the top right of the screen and selected the “comma separated value” inorder to download the data in .csv format as shown in the screenshot below.



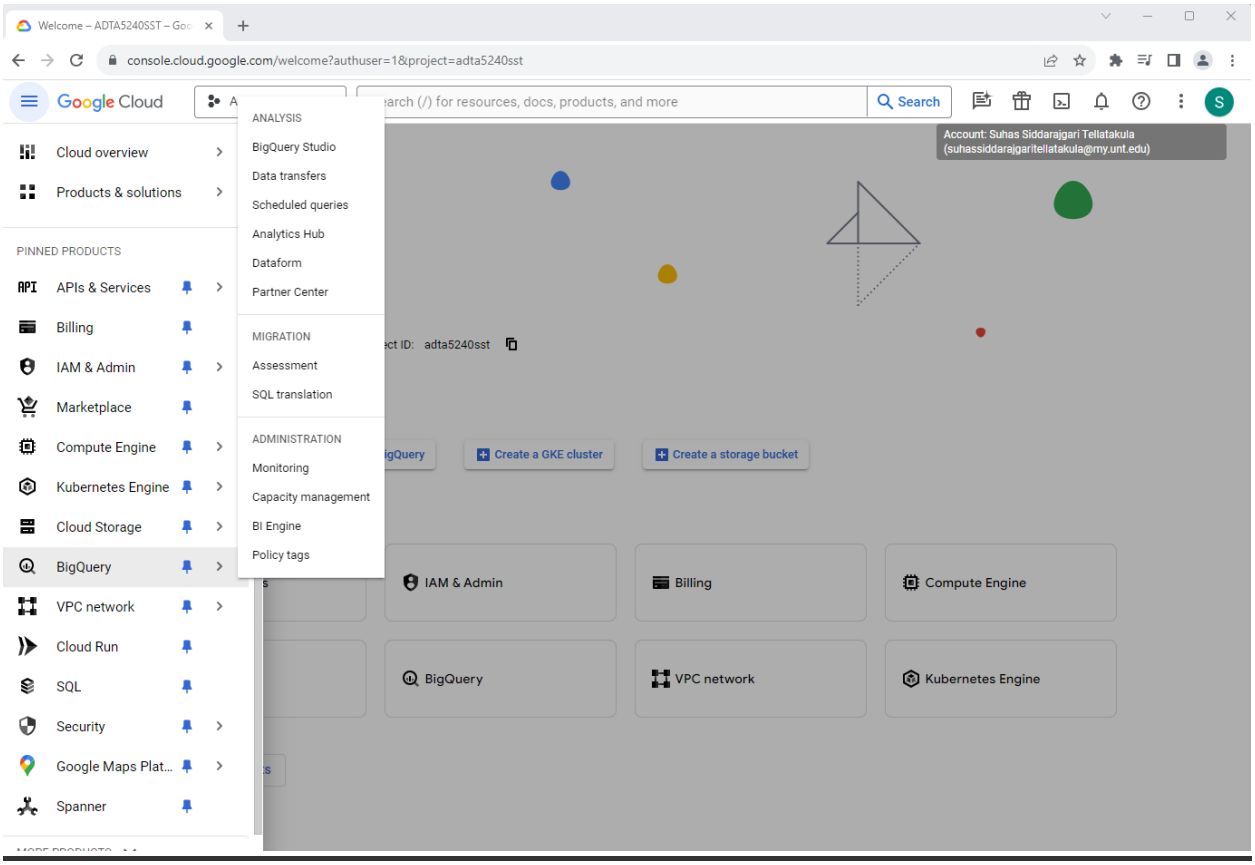
Part 4 – Data Analysis

As a Senior Assistant, I am going to use the Big Query in GCP for data analysis on the cleaned sales records to figure out the 2 results.

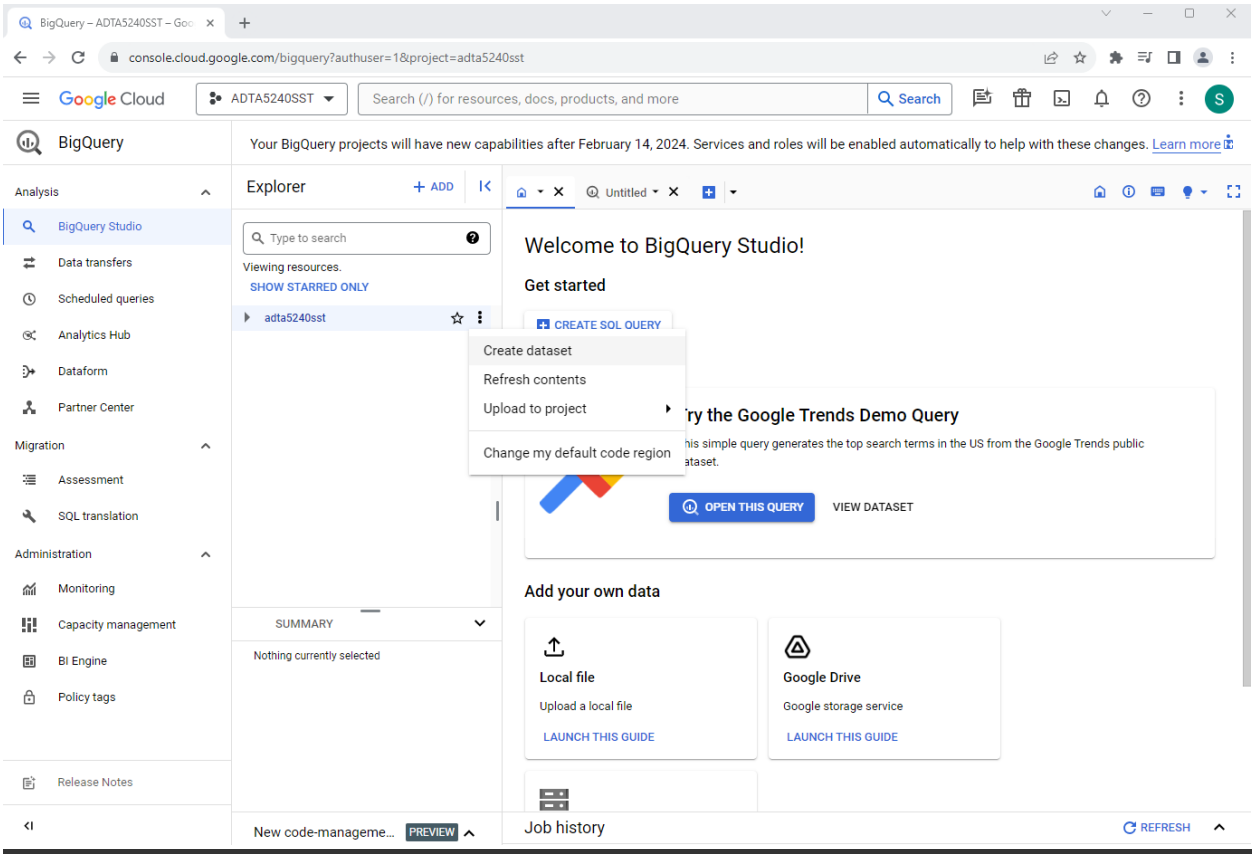
First, I signed into my GCP console which landed me in project home page as shown in below screenshot.



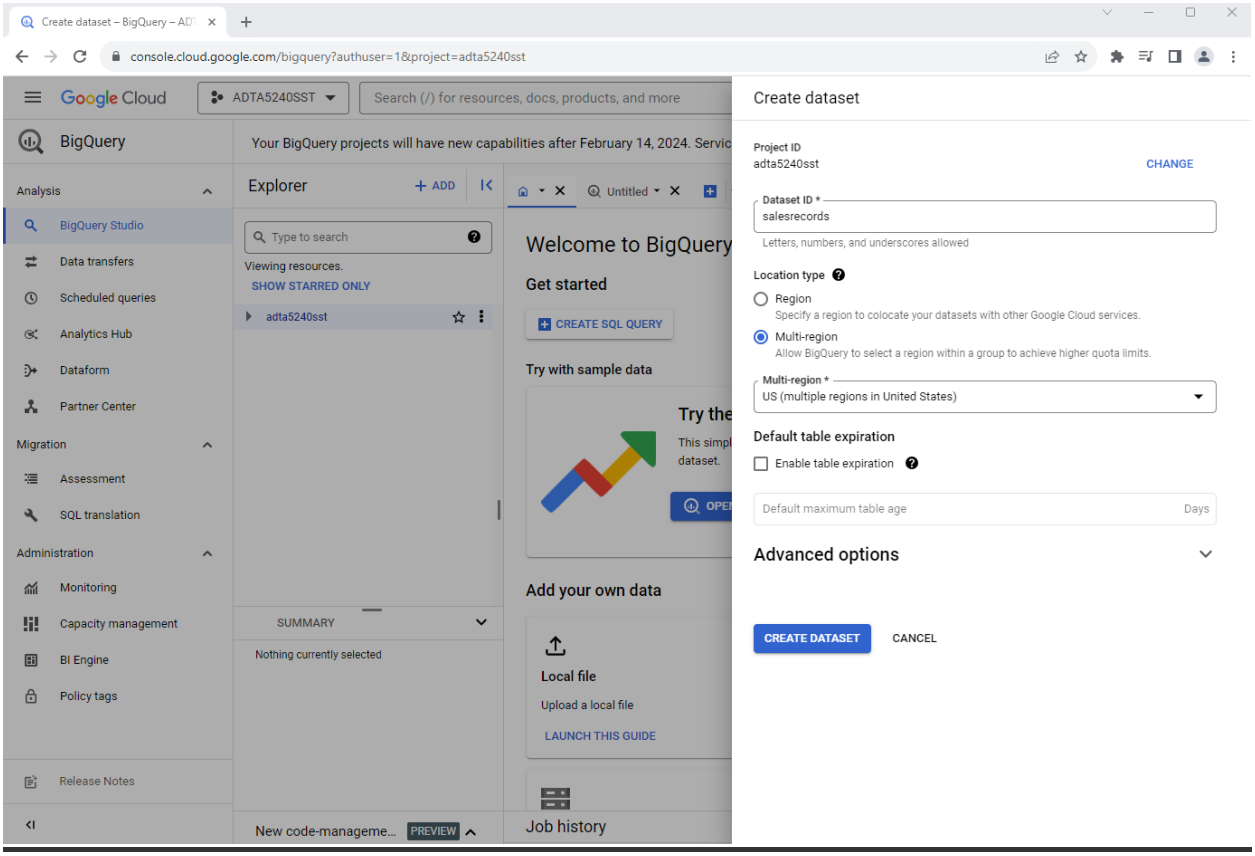
Then, I clicked on the navigation menu on the top left corner and clicked on the Big Query pinned icon as shown in below screenshot.



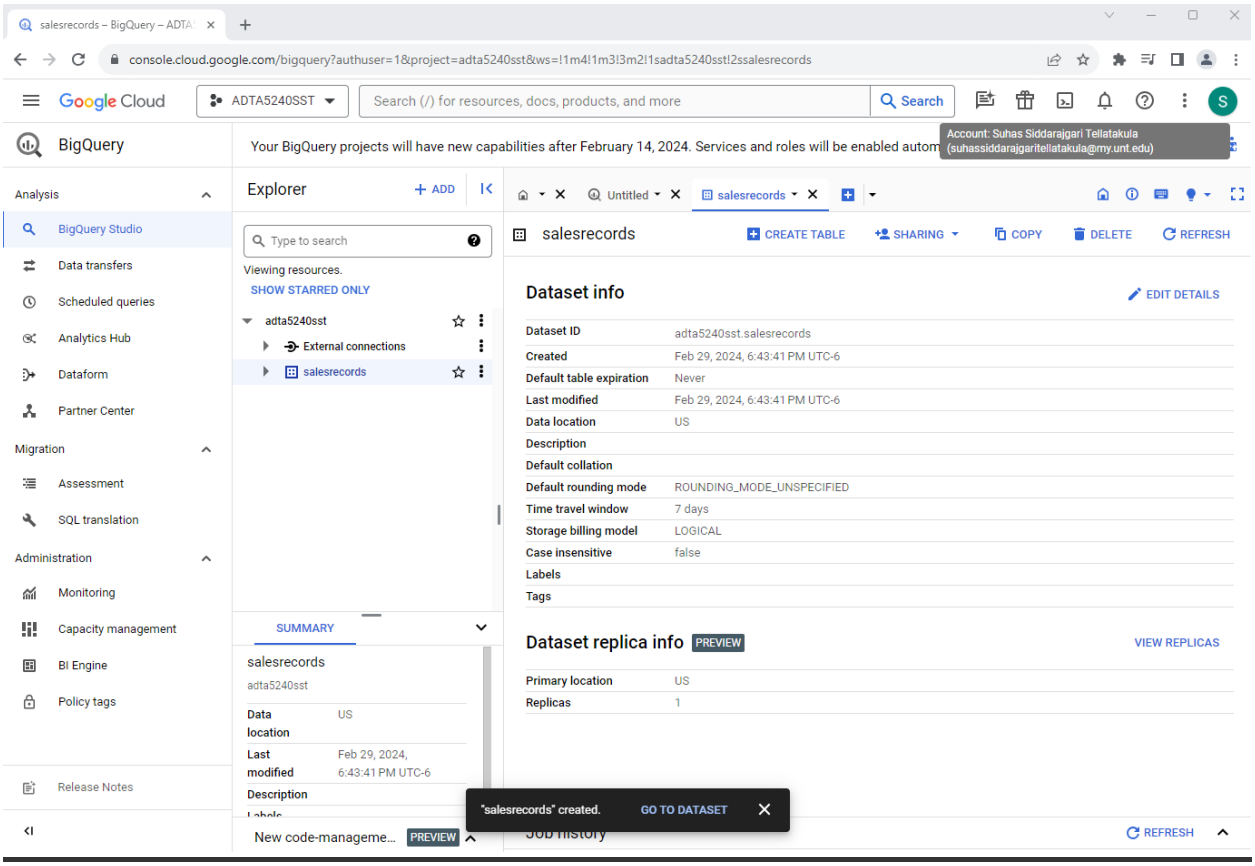
Then, I landed in the Big Query dashboard where we can see my project with 3 dots beside. When clicked on that, we can see an option “create dataset” which allows us to create a new dataset under our project as shown in below screenshot.



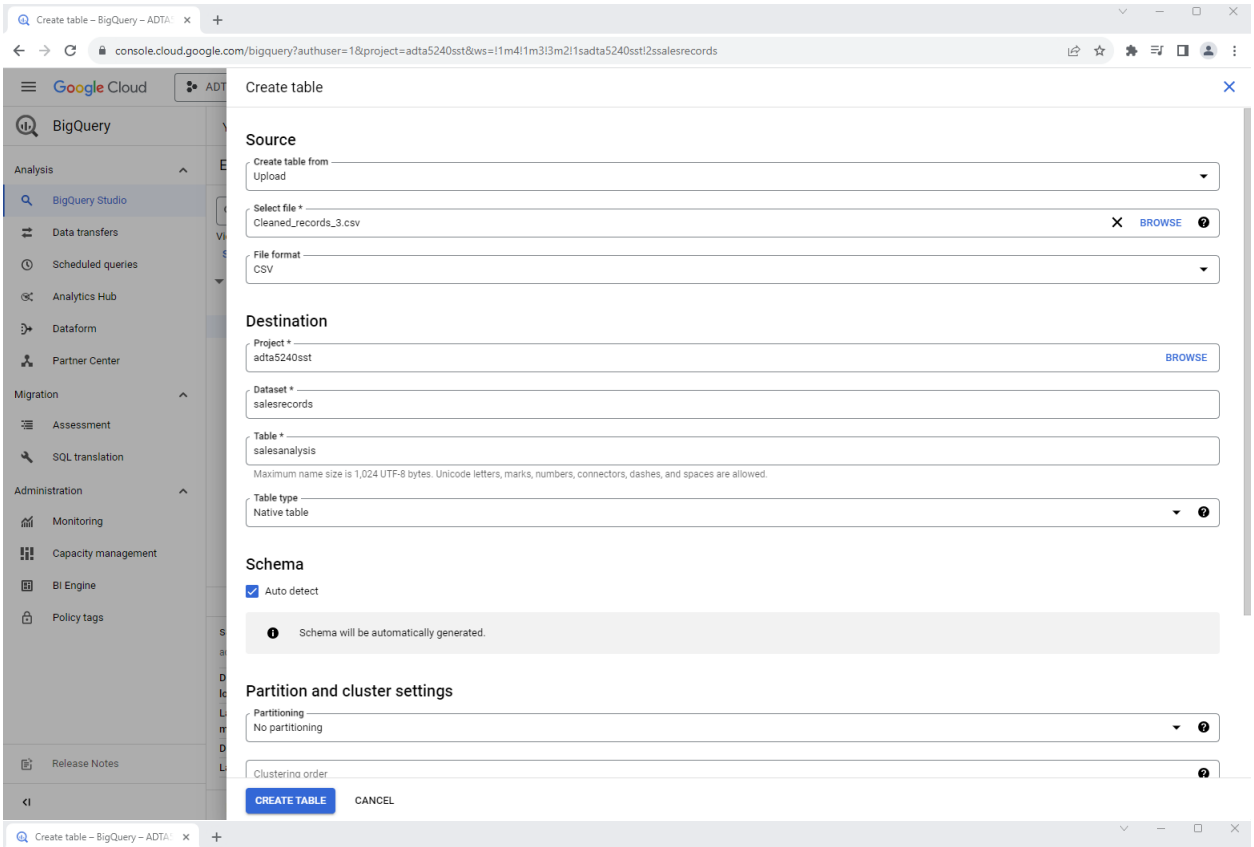
In the create dataset form, I gave “salesrecords” for Dataset ID and chose “US multi-region” and clicked on the blue create dataset button as shown in below screenshot.

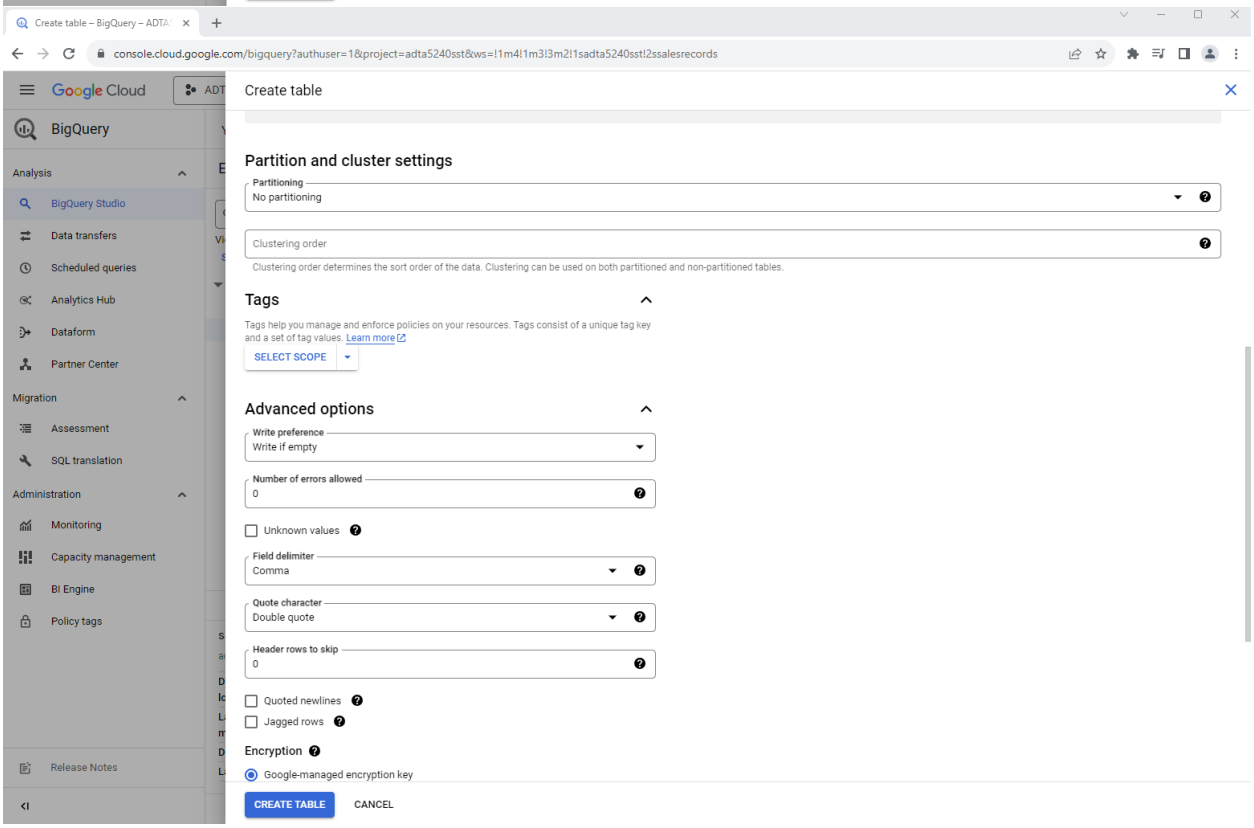


A small message saying dataset is created will pop up. We can see that salesrecords dataset is successfully created and when clicked on it, it takes us to dataset info page with a “Create Table” button which basically allows us to create a new table under our dataset as shown in below screenshot.

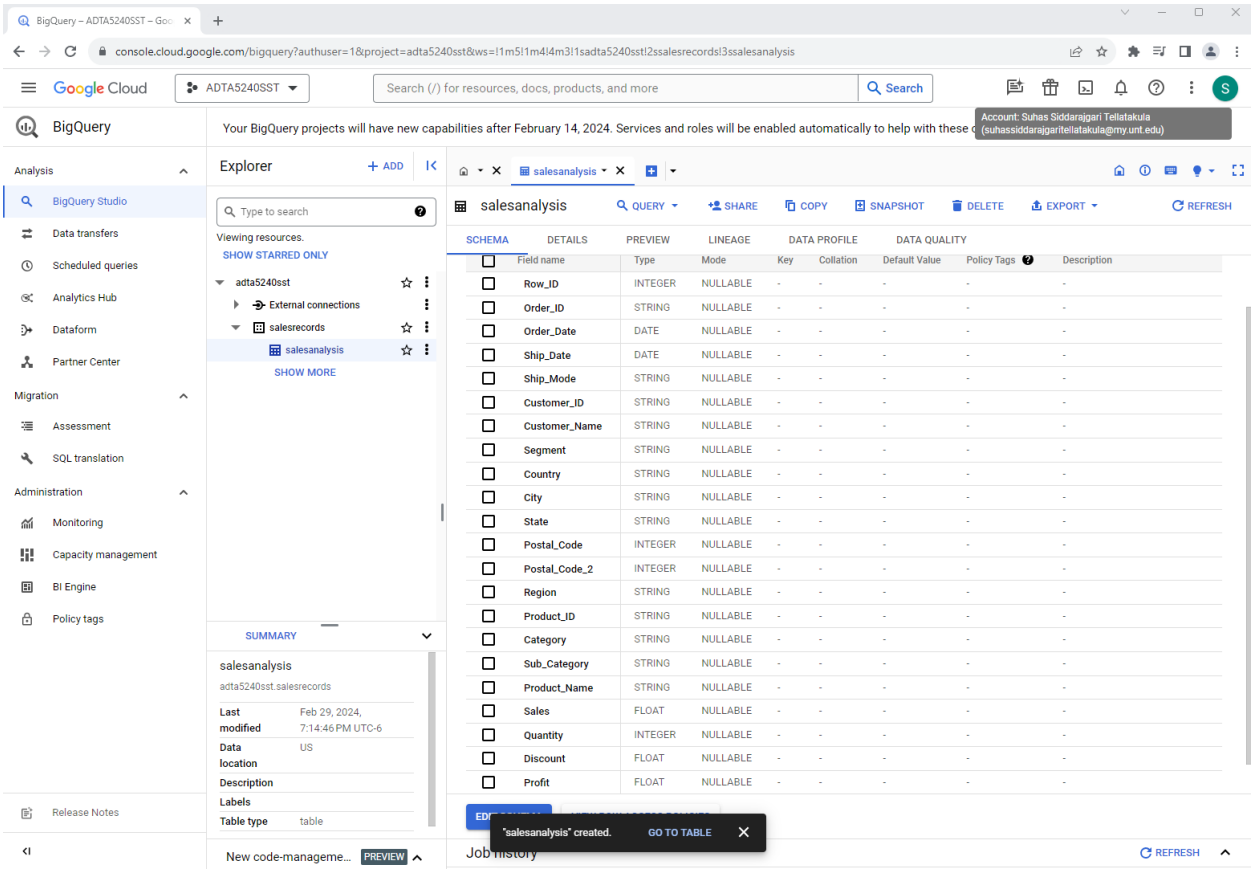


In the create table form, I gave “upload” for Create table from and chose “CSV” for file format and “salesanalysis” as table name. then browsed and selected the Cleaned_records_3.csv” file from my local. I chose Auto detect option for Schema. Then under advanced options, I chose comma for Field delimiter and left the rest as default and finally clicked on the blue create table button as shown in below 2 screenshots.

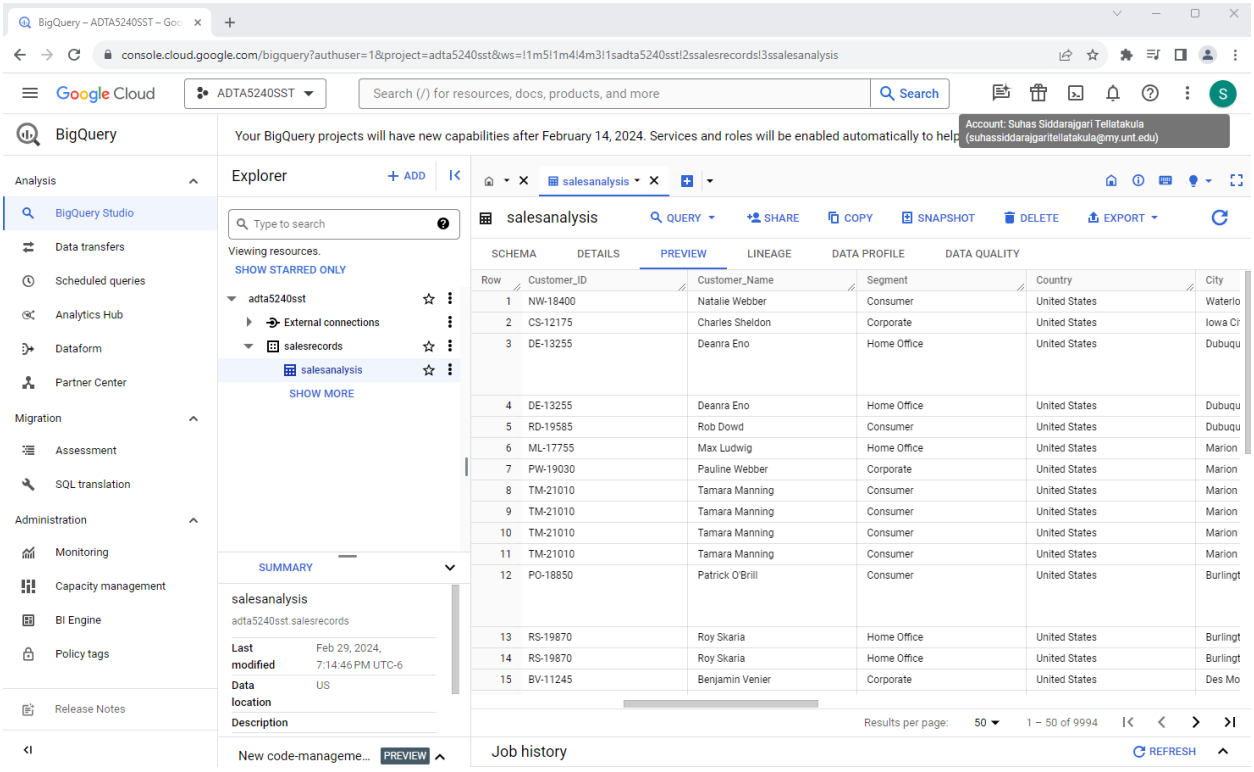

Create table – BigQuery – ADTA
console.cloud.google.com/bigquery?authuser=1&project=adta5240sst&ws=!1m4!1m3!3m2!1sadta5240sst!2ssalesrecords
Google Cloud
BigQuery
Analysis
BigQuery Studio
Data transfers
Scheduled queries
Analytics Hub
Dataform
Partner Center
Migration
Assessment
SQL translation
Administration
Monitoring
Capacity management
BI Engine
Policy tags
Release Notes
Create table
Source
Create table from
Upload
Select file *
Cleaned_records_3.csv
BROWSE
File format
CSV
Destination
Project *
adta5240sst
BROWSE
Dataset *
salesrecords
Table *
salesanalysis
Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.
Table type
Native table
Schema
Auto detect
Schema will be automatically generated.
Partition and cluster settings
Partitioning
No partitioning
Clustering order
CREATE TABLE
CANCEL


Create table – BigQuery – ADTA
console.cloud.google.com/bigquery?authuser=1&project=adta5240sst&ws=!1m4!1m3!3m2!1sadta5240sst!2ssalesrecords
Google Cloud
BigQuery
Analysis
BigQuery Studio
Data transfers
Scheduled queries
Analytics Hub
Dataform
Partner Center
Migration
Assessment
SQL translation
Administration
Monitoring
Capacity management
BI Engine
Policy tags
Release Notes
Create table
Partition and cluster settings
Partitioning
No partitioning
Clustering order
Clustering order determines the sort order of the data. Clustering can be used on both partitioned and non-partitioned tables.
Tags
Tags help you manage and enforce policies on your resources. Tags consist of a unique tag key and a set of tag values. Learn more
SELECT SCOPE
Advanced options
Write preference
Write if empty
Number of errors allowed
0
Unknown values
Field delimiter
Comma
Quote character
Double quote
Header rows to skip
0
Quoted newlines
Jagged rows
Encryption
Google-managed encryption key
CREATE TABLE
CANCEL

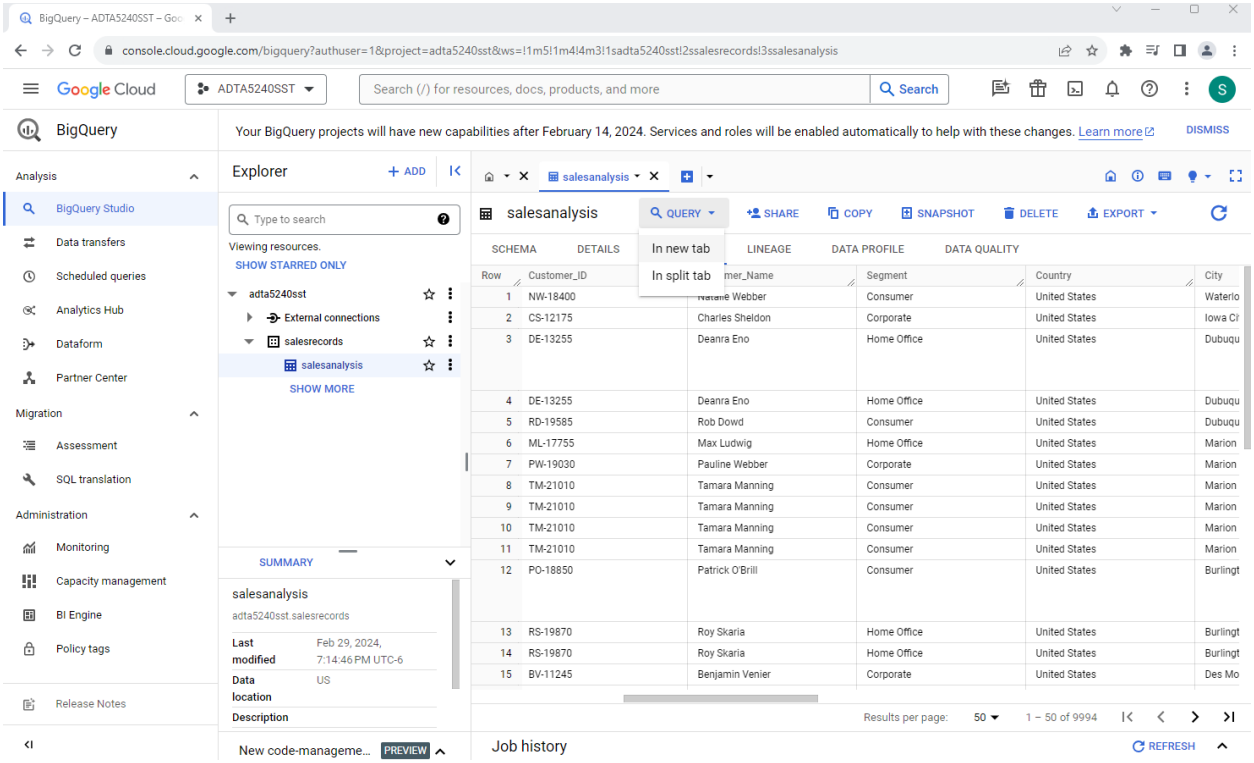
A small message saying salesanalysis created will pop up. We can see that salesanalysis table is successfully created under salesrecords dataset and when clicked on it, it takes us to salesanalysis schemas page which shows us the schema of table as shown in below screenshot.



Then if we click on the preview tab which shows the data inside this table as shown in below screenshot.



In order to write a query in a new tab, I clicked on the query blue button and chose in new tab as shown in below screenshot.



To extract the top 5 states with the greatest number of customers, I entered the following query:

"SELECT State, COUNT(distinct Customer_ID) AS NumCustomers

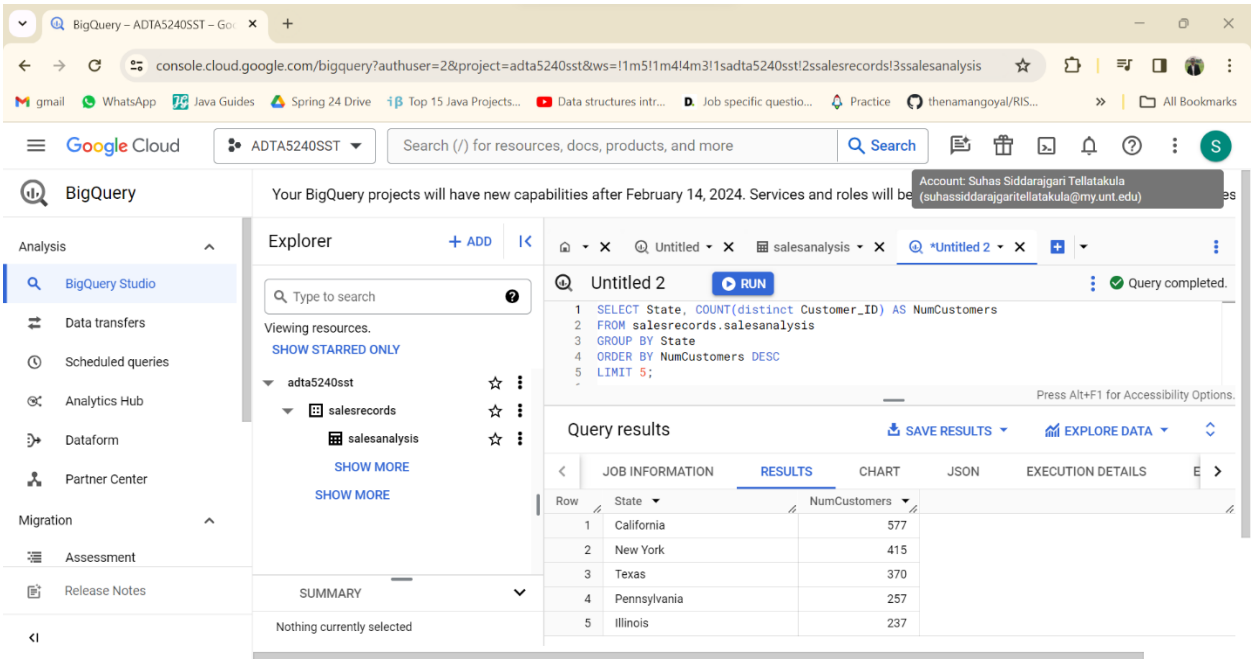
FROM salesrecords.salesanalysis

GROUP BY State

ORDER BY NumCustomers DESC

LIMIT 5;"

Which basically selects the state and frequency of every unique customer_ID in the salesanalysis table grouping state wise and ordering the results based on the frequency value in descending order and limiting the results to only 5 as shown in below screenshot.



In the above screenshot, we can see that the query ran successfully and gave the following results showing that the California state has the top customers with 577 customers, followed by New York with 415, Texas with 370, then Pennsylvania with 257 and finally at 5th position is Washington with 237 customers.

Row	State	NumCustomers
1	California	577
2	New York	415
3	Texas	370
4	Pennsylvania	257
5	Washington	237

To extract the top 10 city zip codes with the highest sales value, I entered the following query:

"SELECT Postal_Code, SUM (Sales) AS TotalSales

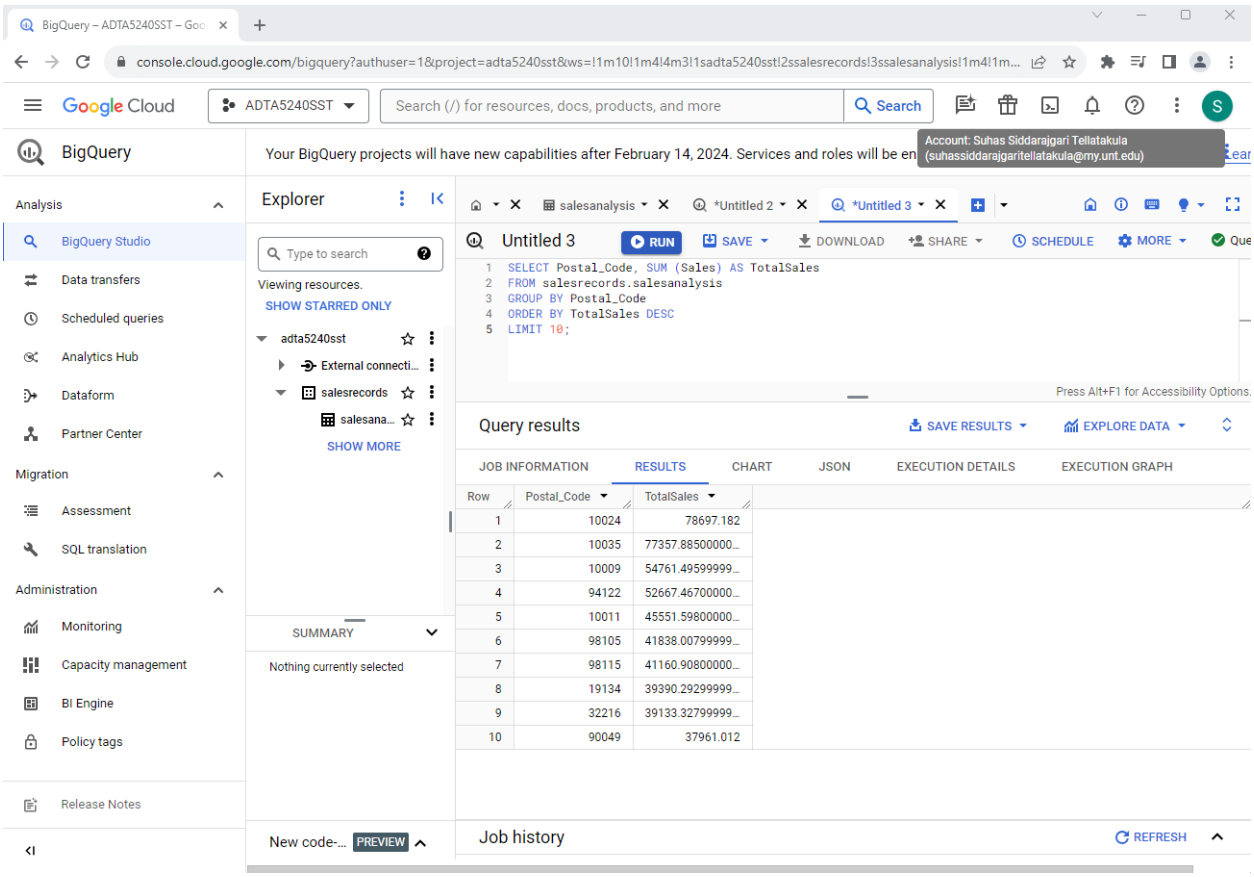
FROM salesrecords.salesanalysis

GROUP BY Postal_Code

ORDER BY TotalSales DESC

LIMIT 10;"

Which basically selects the zip codes and total values of Sales in the salesanalysis table grouping city postal code wise and ordering the results based on the total values value in descending order and limiting the results to the first 10 as shown in below screenshot.



In the above screenshot, we can see that the query ran successfully and gave the following results showing that the city with postal code 10024 has the highest sales with value of 78697.182, followed by 10035 city with almost 77358, and so on with 10th position for city with 90049 with sales value of approximately 37961.

Row	Postal_Code	TotalSales
1	10024	78697.182
2	10035	77357.885000000009
3	10009	54761.495999999963
4	94122	52667.467000000011
5	10011	45551.598000000013
6	98105	41838.00799999998
7	98115	41160.90800000001
8	19134	39390.292999999976
9	32216	39133.327999999994
10	90049	37961.012