

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.

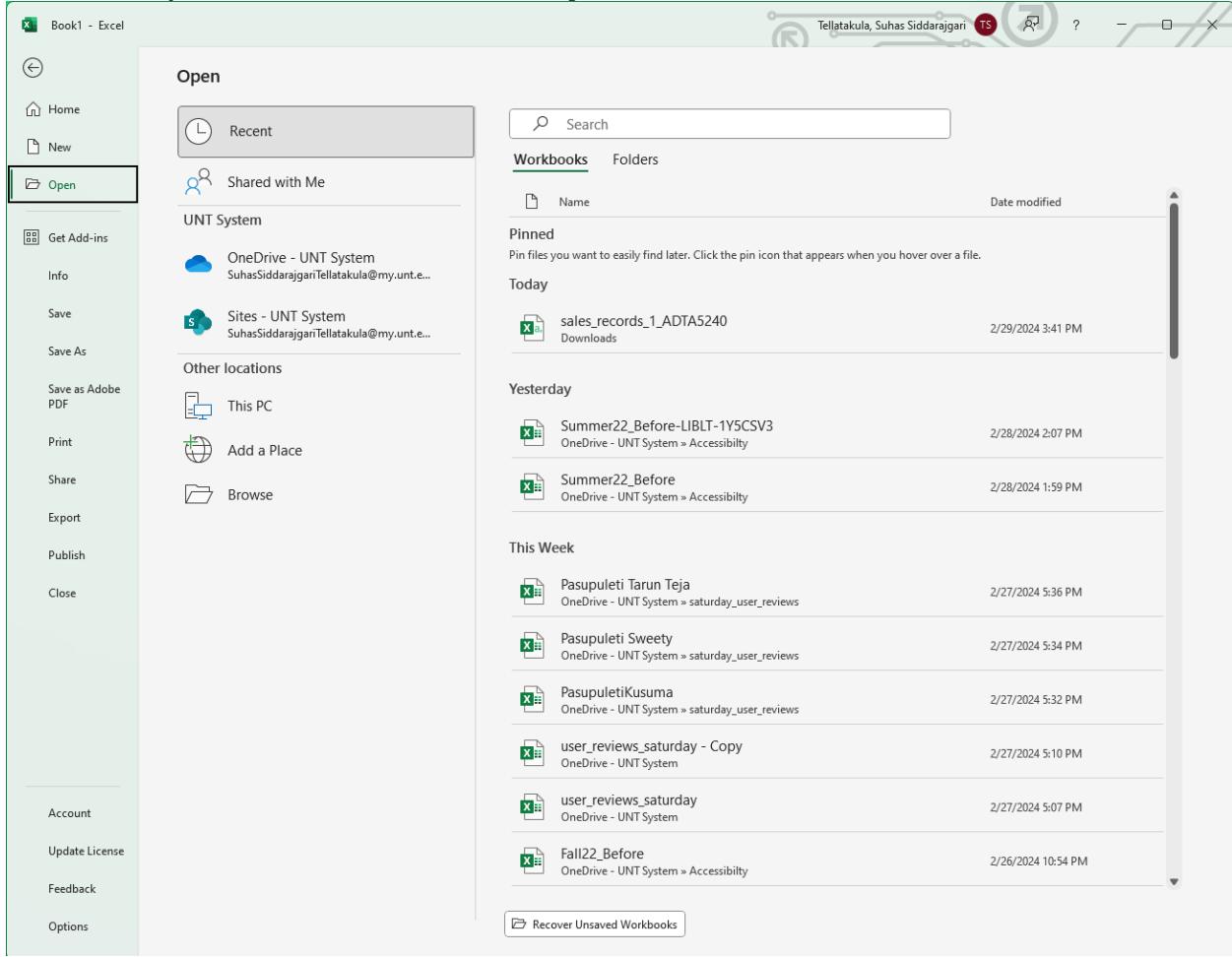
Conclusion:

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.

RowID	OrderID	Order Date	Ship Date	Ship Mode	Customer	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit
1	CA-2016-1 #####	6/9/2014	6/10/2014	Second	Cl-G-12520	Claire Gut	Consumer	United Sta	Henderson	Kentucky	42420	South	FUR-BO-1C Furniture	Bookcases	Bush Somerset Collection Bookcase	261.96	2	0	41.9136
2	CA-2016-1 #####	6/9/2014	6/10/2014	Second	Cl-CG-12520	Claire Gut	Consumer	United Sta	Henderson	Kentucky	42420	South	FUR-CH-1C Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking	731.94	3	0	219.582
3	CA-2016-1 #####	6/9/2014	6/10/2014	Second	Cl-DV-13045	Darrin Van	Corporate	United Sta	Los Angeles	California	90036	West	OFF-LA-10 Office Supy	Labels	Self-Adhesive Address Labels for Typewri	14.62	2	0	6.8714
4	US-2015-1 #####	6/9/2014	6/10/2014	Standard	CS-20335	Sean O'Do	Consumer	United Sta	Fort Lauderdale	Florida	33311	South	FUR-TA-10 Furniture	Tables	Bretford CR4500 Series Slim Rectangular	957.5775	5	0.45	-383.031
5	US-2015-1 #####	6/9/2014	6/10/2014	Standard	CS-20335	Sean O'Do	Consumer	United Sta	Fort Lauderdale	Florida	33311	South	OFF-ST-10 Furniture	Storage	Eldon Fold N Roll Cart System	22.368	2	0.2	4.5164
6	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032	West	FUR-FU-1C Furniture	Furnishing	Elton Expressions Wood and Plastic Des	48.86	7	0	14.1694
7	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032	West	OFF-AR-1C Office Supy	Art	Newell 322	7.28	4	0	1.9656
8	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032	West	TEC-PH-1C Technolog	Phones	Mitel 5320 IP Phone VoIP phone	907.152	6	0.2	90.7152
9	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032	West	OFF-BI-10 Office Supy	Binders	DXL Angle-View Binders with Locking Rin	18.504	3	0.2	5.7825
10	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032	West	OFF-AP-10 Office Supy	Appliance	Belkin F5C206VTEL 6 Outlet Surge	114.9	5	0	34.47
11	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032	West	FUR-TA-10 Furniture	Tables	Chromcraft Rectangular Conference Tab	1706.184	9	0.2	85.3092
12	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032	West	TEC-PH-1C Technolog	Phones	Konfetti 25 Conference phone - Charco	911.424	4	0.2	68.3568
13	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032	West	OFF-PA-10 Office Supy	Art	Xerox 1967	15.552	3	0.2	5.4432
14	CA-2017-1 #####	6/9/2014	6/10/2014	Standard	CAA-10480	Andrew All	Consumer	United Sta	Concord	North Carr	28027	South	OFF-PA-10 Office Supy	Binders	Fellowes PB200 Plastic Comb Binding M	407.976	3	0.2	132.5922
15	CA-2016-1 #####	6/9/2014	6/10/2014	Standard	CM-15070	Irene Madc	Consumer	United Sta	Seattle	Washington	98103	West	OFF-BI-10 Office Supy	Binders	Holmes Replacement Filter for HEPA Air	68.81	5	0.8	-123.858
16	US-2015-1 #####	6/9/2014	6/10/2014	Standard	CHP-14815	Harold Paw	Home Offic	United Sta	Fort Worth	Texas	76106	Central	OFF-AP-10 Office Supy	Appliance	Storex DuraTech Recyclable Plastic Froste	2.544	3	0.8	-3.816
17	US-2015-1 #####	6/9/2014	6/10/2014	Standard	CHP-14815	Harold Paw	Home Offic	United Sta	Fort Worth	Texas	76106	Central	OFF-BI-10 Office Supy	Binders	Stur-D-O-Stor Shelving, Vertical 5-Shelf, 72	665.88	6	0	13.3176
18	CA-2014-1 #####	6/9/2014	6/10/2014	Standard	CPK-19075	Pete Kitz	Consumer	United Sta	Madison	Wisconsin	53711	Central	OFF-ST-10 Office Supy	Storage	Fellowes Super Stor/Drawer	55.5	2	0	9.99
19	CA-2014-1 #####	6/9/2014	6/10/2014	Standard	CPK-19075	Pete Kitz	Consumer	United Sta	West Jord	Utah	84084	West	OFF-ST-10 Office Supy	Storage	Fellowes Super Stor/Drawer	8.56	2	0	2.4824
20	CA-2014-1 #####	6/9/2014	6/10/2014	Second	Cl-ZD-21925	Zuschuss	Consumer	United Sta	San Franci	California	94109	West	OFF-AR-1C Office Supy	Art	Wilson Newell 341	213.48	3	0.2	16.0111
21	CA-2014-1 #####	6/9/2014	6/10/2014	Second	Cl-ZD-21925	Zuschuss	Consumer	United Sta	San Franci	California	94109	West	TEC-PH-1C Technolog	Phones	Wilson Jones Hanging View Binder, White	22.72	4	0.2	7.384
22	CA-2014-1 #####	6/9/2014	6/10/2014	Second	Cl-ZD-21925	Zuschuss	Consumer	United Sta	San Franci	California	94109	West	OFF-BI-10 Office Supy	Binders	Acco Six-Outlet Power Strip, 4' Cord Len	19.46	7	0	5.0596
23	CA-2016-1 #####	6/9/2014	6/10/2014	Standard	CKB-16585	Ken Black	Corporate	United Sta	Fremont	Nebraska	68025	Central	OFF-AR-1C Office Supy	Art	Global Deluxe Stacking Chair, Gray	60.34	7	0	15.6884
24	CA-2016-1 #####	6/9/2014	6/10/2014	Standard	CKB-16585	Ken Black	Corporate	United Sta	Fremont	Nebraska	68025	Central	OFF-FU-1C Office Supy	Furniture	Howard Miller 13-3/4" Diameter Brushed	71.372	2	0.3	-1.0196
25	CA-2017-1 #####	6/9/2014	6/10/2014	Second	Cl-SF-20065	Sandra Fla	Consumer	United Sta	Philadelphia	Pennsylvn	19140	East	FUR-CH-1C Furniture	Chairs	Bretford CR4500 Series Slim Rectangular	1044.63	3	0	240.2649
26	CA-2016-1 #####	6/9/2014	6/10/2014	Standard	Cl-EH-13945	Eric Hoffm	Consumer	United Sta	Los Angeles	California	90049	West	OFF-BI-10 Office Supy	Binders	Wilson Jones Active Use Binders	11.648	2	0.2	4.2224
27	CA-2016-1 #####	6/9/2014	6/10/2014	Second	Cl-EH-13945	Eric Hoffm	Consumer	United Sta	Los Angeles	California	90049	West	TEC-AC-1C Technolog	Accessori	Imation 8GB Mini TravelDrive USB 2.0 Fl	90.57	3	0	11.7741
28	CA-2016-1 #####	6/9/2014	6/10/2014	Standard	CTB-21520	Tracy Blum	Consumer	United Sta	Philadelphia	Pennsylvn	91940	West	OFF-BI-10 Office Supy	Binders	Riverside Palais Royal Lawyers Bookcase	3083.43	7	0.5	-1665.05
29	US-2015-1 #####	6/9/2014	6/10/2014	Standard	CTB-21520	Tracy Blum	Consumer	United Sta	Philadelphia	Pennsylvn	91940	West	OFF-BI-10 Office Supy	Binders	Avery Recycled Flexi-View Covers for Bir	9.618	2	0.7	-7.0532
30	US-2015-1 #####	6/9/2014	6/10/2014	Standard	CTB-21520	Tracy Blum	Consumer	United Sta	Philadelphia	Pennsylvn	91940	West	OFF-FU-1C Office Supy	Furniture	Howard Miller 13-3/4" Diameter Brushed	124.2	3	0.2	15.525
31	US-2015-1 #####	6/9/2014	6/10/2014	Standard	CTB-21520	Tracy Blum	Consumer	United Sta	Philadelphia	Pennsylvn	91940	West	OFF-EN-1C Office Supy	Envelopes	Poly String Tie Envelopes	3.264	2	0.2	1.1016
32	US-2015-1 #####	6/9/2014	6/10/2014	Standard	CTB-21520	Tracy Blum	Consumer	United Sta	Philadelphia	Pennsylvn	91940	West	OFF-AR-1C Office Supy	Art	BOSTON Model 1800 Electric Pencil Sha	86.304	6	0.2	9.7092
33	US-2015-1 #####	6/9/2014	6/10/2014	Standard	CTB-21520	Tracy Blum	Consumer	United Sta	Philadelphia	Pennsylvn	91940	West	OFF-BI-10 Office Supy	Binders	Acco Pressboard Covers with Storage Hd	6.858	6	0.7	-5.715
34	US-2015-1 #####	6/9/2014	6/10/2014	Standard	CTB-21520	Tracy Blum	Consumer	United Sta	Philadelphia	Pennsylvn	91940	East	OFF-AR-1C Office Supy	Art	Lumber Crayons	15.76	2	0.2	3.546
35	US-2017-1 #####	6/9/2014	6/10/2014	Second	Cl-MA-17560	Matt Abelin	Home Offic	United Sta	Houston	Texas	77095	Central	OFF-PA-10 Office Supy	Paper	Easy-staple paper	29.472	3	0.2	9.9468
36	CA-2016-1 #####	6/9/2014	6/10/2014	First Class	GH-14485	Gene Hale	Corporate	United Sta	Richardson	Texas	75080	Central	TEC-PH-1C Technolog	Phones	GE 30524E4	1097.544	7	0.2	123.4737
37	CA-2016-1 #####	6/9/2014	6/10/2014	First Class	GH-14485	Gene Hale	Corporate	United Sta	Richardson	Texas	75080	Central	FUR-CH-1C Furniture	Furnishing	Electrix Architect's Clamp-On Swing Arm	390.92	5	0.6	-147.963
38	CA-2015-1 #####	6/9/2014	6/10/2014	Standard	CSN-20710	Steve Ngai	Home Offic	United Sta	Houston	Texas	77041	Central	OFF-EN-1C Office Supy	Envelopes	#10 4 1/8" x 9 1/2" Premium Diagonal Se	113.328	9	0.2	35.415
39	CA-2015-1 #####	6/9/2014	6/10/2014	Standard	CSN-20710	Steve Ngai	Home Offic	United Sta	Houston	Texas	77041	Central	FUR-BO-1C Furniture	Bookcases	Atlantic Metals Mobile 3-Shelf Bookcase	532.3992	3	0.32	-46.9764
40	CA-2015-1 #####	6/9/2014	6/10/2014	Standard	CSN-20710	Steve Ngai	Home Offic	United Sta	Houston	Texas	77041	Central	FUR-CH-1C Furniture	Chairs	Global Fabric Manager's Chair, Dark Gra	212.058	3	0.3	-15.147

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.

The screenshot shows a Microsoft Excel spreadsheet titled "sales_records_1_ATDA5240". The spreadsheet contains data from rows 1 to 41, spanning columns A through U. The data includes columns for Row ID, Order ID, Order Date, Ship Date, Ship Mode, Customer Segment, Country, City, State, Postal Code, Region, Product ID, Category, Sub-Category, Product Name, Sales, Quantity, Discount, and Profit. A "Find and Replace" dialog box is overlaid on the spreadsheet, with the "Find" tab selected. The "Find what:" field contains the text "Standard C". The "Replace with:" field is empty. Below the input fields are three buttons: "Find All", "Find Next", and "Close". The background of the spreadsheet shows a grid of data corresponding to the columns and rows defined above.

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.

The screenshot shows a Microsoft Excel spreadsheet titled "sales_records_1_ADTA..". The "Replace" dialog box is open in the foreground, centered over the data. The dialog has tabs for "Find" and "Replace". The "Replace" tab is selected. In the "Find what:" field, there is placeholder text "Find what: []". In the "Replace with:" field, there is placeholder text "Replace with: []". Below these fields are buttons for "Replace All", "Replace", "Find All", and "Find Next". A "Close" button is also present. The background spreadsheet contains data from rows 1 to 41, with columns labeled A through U. The data includes various product details like Product ID, Category, Sub-Categ, Product Name, Sales, Quantity, Discount, and Profit. The "Replace" dialog is partially obscuring the bottom portion of the spreadsheet.

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.

The screenshot shows a Microsoft Excel spreadsheet titled "sales_records_1_ADTA5240". The spreadsheet contains data for various products, including their names, descriptions, and sales details. A search and replace dialog box is overlaid on the spreadsheet. The "Find what:" field contains a comma (,), and the "Replace with:" field contains a hyphen (-). The "Replace All" button is highlighted, indicating it has been selected. The dialog also includes "Find", "Replace", "Find All", and "Find Next" buttons, along with an "Options >" button.

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Customer	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit
1	CA-2016-1 #####	6/9/2014	6/10/2014	Second Class	Craig Gut	Consumer	United States	Henderson	Kentucky		42420	South	FUR-B0-11 Furniture	Bookcases	Bush Somerset Collection	Bookcase	261.96	2	0	41.9136
2	CA-2016-1 #####	6/9/2014	6/10/2014	Second Class	Craig Gut	Consumer	United States	Henderson	Kentucky		42420	South	FUR-CH-11 Furniture	Chairs	Han Deluxe Fabric Upholstered Stacking	731.94	3	0	219.582	
3	CA-2016-1 #####	6/9/2014	6/10/2014	Second Class	Darin Van	Corporate	United States	Los Angeles	California		90036	West	OFF-FLA-10 Office Suply Labels	Labels	Self-Adhesive Address Labels for Typewri	14.62	2	0	6.8714	
4	US-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CSO-20335	Sean O'Do	Consumer	United States	Ft Lauderdale	Florida	33311	South	FUR-TA-10 Furniture	Tables	Bretford CR4500 Series Slim Rectangula	957.5775	5	0.45	-383.031	
5	US-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CSO-20335	Sean O'Do	Consumer	United States	Ft Lauderdale	Florida	33311	South	OFF-ST-10 Office Suply Storage	Storage	Eldon Fold 'N Roll Cart System	22.368	2	0.2	2.5164	
6	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-1C Furniture	Furnishing	Eldon Expressions Wood and Plastic Des	48.86	7	0	14.1694	
7	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-1C Office Suply Art	Art	Newell 322	7.28	4	0	1.9656	
8	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	TEC-PB200 IP Phone	VoIP phones	Mitel 5320 IP Phone VoIP phone	907.152	6	0.2	90.7152	
9	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	DXL Angle-View Binders with Locking Ring	18.504	3	0.2	5.7825	
10	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	Belkin F5C206VTEL 6 Outlet Surge	114.9	5	0	34.47	
11	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Furniture	Tables	Chromcraft Rectangular Conference Tab	1706.184	9	0.2	85.3092	
12	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Technolo	Phones	Kontrol 25 Conference phone - Charcoal	911.424	4	0.2	68.3568	
13	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Paper	Paper	Xerox 1967	15.552	3	0.2	5.4432	
14	CA-2014-1 6/9/2014 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	Fellowes PB200 Plastic Comb Binding Ma	407.976	3	0.2	132.5922	
15	US-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Appliances	Appliances	Belkin F5C206VTEL 6 Outlet Surge	68.81	5	0.8	-123.858	
16	US-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	Holmes Replacement Filter for HEPA Air	2.544	3	0.8	-3.816	
17	CA-2014-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	Stor-D-Stor Recycled Plastic Froste	665.88	6	0	13.3176	
18	CA-2014-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Storage	Storage	Stur-D-Stor Shelving, Vertical 5-Shelf, 72	55.5	2	0	9.99	
19	CA-2014-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Art	Art	Fellowes Super Stor Drawer	8.56	2	0	2.4824	
20	CA-2014-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Technolo	Phones	Cisco SPA501G IP Phone	213.48	3	0.2	16.0111	
21	CA-2014-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	Wilson Jones Hanging View Binder, White	22.72	4	0.2	7.384	
22	CA-2014-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Art	Art	Newell 318	19.46	7	0	5.0596	
23	CA-2014-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Appliances	Appliances	Acco Six-Outlet Power Strip, 4' Cord Leng	60.34	7	0	15.6884	
24	US-2017-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Furniture	Chairs	Global Deluxe Stacking Chair, Gray	71.372	2	0.3	-1.0196	
25	CA-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Furniture	Tables	Brettford CR4500 Series Slim Rectangula	1044.63	3	0	240.2649	
26	CA-2016-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	Wilson Jones Active Use Binders	11.648	2	0.2	4.2224	
27	CA-2016-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Technolo	Accessorie	Imatron 8GB Mini TravelDrive USB 2.0 Fl	90.57	3	0	11.7741	
28	CA-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Furniture	Bookcases	Riverside Palais Royal Lawyers Bookcase	3083.43	7	0.5	-1665.05	
29	CA-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	Avery Recycled Flexi-View Covers for Bin	9.618	2	0.7	-7.0532	
30	CA-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Furniture	Furnishing	Howard Miller 13-3/4" Diameter Brushe	124.2	3	0.2	15.525	
31	US-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Envelopes	Envelopes	PolyString Tie Envelopes	3.264	2	0.2	1.1016	
32	CA-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Art	Art	BOSTON Model 1800 Electric Pencil Sh	86.304	6	0.2	9.7092	
33	US-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	Acco Pressboard Covers with Storage Ho	6.658	6	0.7	-5.715	
34	US-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Art	Art	Lumber Crayon	15.76	2	0.2	3.546	
35	CA-2017-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	Easy-staple paper	29.472	3	0.2	9.9468	
36	CA-2016-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Appliances	Appliances	GE 30524EE4	1097.544	7	0.2	123.4737	
37	CA-2016-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Furniture	Tables	FUR-FU-11 Furniture	190.92	5	0.6	-147.963	
38	CA-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Furniture	Furnishing	Electrix Architect's Clamp-On Swing Arm	113.328	9	0.2	35.415	
39	CA-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	Atlantic Metals Mobile 3-Shelf Bookcas	532.3992	3	0.32	-46.9764	
40	CA-2015-1 #####	6/9/2014	6/10/2014	Standard Class	CBH-11710	Brosina Hc	Consumer	United States	Los Angeles	California	90032	West	Office Supy Binders	Binders	Global Fabric Manager's Chair, Dark Gra	212.058	3	0.3	-15.147	

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

The screenshot shows a Microsoft Excel spreadsheet titled "sales_records_1_ADTA...". The spreadsheet contains data from rows 1 to 41, with columns A through U. The data includes various product details such as Product ID, Category, Sub-Catag, Product Name, Sales, Quantity, Discount, and Profit. A "Find and Replace" dialog box is open in the foreground, centered over the data. The dialog box has the title "Microsoft Excel" and the message "All done. We made 3120 replacements." Below the message are "OK" and "Options >" buttons. At the bottom of the dialog are buttons for "Replace All", "Replace", "Find All", "Find Next", and "Close". The status bar at the bottom of the Excel window shows "Ready" and "Accessibility: Unavailable".

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Customer Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Catag	Product Name	Sales	Quantity	Discount	Profit
1	CA-2016-1	6/9/2014		Second	Cl-G-12520	Claire Gut	Consumer	United Sta	Henderson	Kentucky	42420 South	FUR-B0-11 Furniture	Bookcases	Bush Somerset Collection	Bookcase	261.96	2	0	41.9136
2	CA-2016-1	6/9/2014		Second	Cl-G-12520	Claire Gut	Consumer	United Sta	Henderson	Kentucky	42420 South	FUR-CH-11 Furniture	Chairs	Hon Deluxe Fabric Upholstered	Stacking	731.94	3	0	219.582
3	CA-2016-1	6/9/2014		Second	Cl-DV-13045	Darrin Van	Corporate	United Sta	Los Angeles	California	90036 West	OFF-LA-10 Office	Sup Labels	Self-Adhesive Address	Labels for Typewri	14.62	2	0	6.6714
4	US-2015-1	6/9/2014		Standard	CSO-20335	Sean O'Do	Consumer	United Sta	Fort Lauderdale	Florida	33311 South	FUR-TA-10 Furniture	Tables	Bretford CR4500 Series	Slim Rectangula	95.5775	5	0.45	-383.031
5	US-2015-1	6/9/2014		Standard	CSO-20335	Sean O'Do	Consumer	United Sta	Fort Lauderdale	Florida	33311 South	OFF-ST-10 Office	Sup Storage	Eldon Fold N Roll	Cart System	22.368	2	0.2	2.5164
6	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	FUR-FU-11 Furniture	Furnishing	Eldon Expressions	Wood and Plastic Des	48.86	7	0	14.1694
7	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	OFF-AR-1C Office	Sup Art	Newell 322		7.28	4	0	1.9656
8	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	TEC-PH-1C Technology	Phones	Mitel 5320 IP	Phone VoIP phone	907.152	6	0.2	90.7152
9	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	OFF-BI-10 Office	Sup Binders	DXL Angle-View	Binders with Locking Rim	18.504	3	0.2	5.7825
10	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Appliances	Belkin F5C206VTEL6	Outlet Surge	114.9	5	0	34.47
11	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Tables	Chromcraft	Rectangular Conference	1706.184	9	0.2	85.3092
12	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Technolog	Phones	Konfet	250 Conference phone - Charco	911.424	4	0.2	68.3568
13	CA-2017-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Paper	Xerox 1967		15.552	3	0.2	5.4432
14	CA-2016-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Binders	Fitelows PB200	Plastic Comb Binding M	407.976	3	0.2	132.5922
15	US-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Appliances	Holmes	Replacement Filter for HEPA Air	68.81	5	0.8	-123.858
16	US-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Binders	Storx DuraTech	Recycled Plastic Frost	2.544	3	0.8	-3.816
17	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Storage	Stur-D-Stor	Shelving- Vertical 5-Shelf	655.88	6	0	13.3176
18	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Storage	Fellowes Super Stor	/Drawer	55.5	2	0	9.99
19	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Art	Newell 341		8.56	2	0	2.4824
20	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Technolog	Phones	Cisco SPA5010	IP Phone	213.48	3	0.2	16.0111
21	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Binders	Wilson Jones	Hanging View Binder- Whit	22.72	4	0.2	7.384
22	CA-2014-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Art	Newell 318		19.46	7	0	5.0596
23	CA-2016-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Appliances	Acco Six-Outlet	Power Strip- 4' Cord Len	60.34	7	0	15.6884
24	CA-2017-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Furniture	Chairs	Global Deluxe	Stacking Chair- Gray	71.372	2	0.3	-1.0196
25	CA-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Furniture	Tables	Bretford CR4500 Series	Slim Rectangula	1044.63	3	0	240.2649
26	CA-2016-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Binders	Wilson Jones	Active Use Binders	116.68	2	0.2	4.2224
27	CA-2016-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Technolog	Accessorie	Imation 8GB	Mini TravelDrive USB 2.0 Flc	90.57	3	0	11.7741
28	US-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Furniture	Bookcases	Riverside Palais Royal	Lawyers Bookcas	3083.43	7	0.5	-1665.05
29	US-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Binders	Avery Recycled	Flexi-View Covers for Bin	9.618	2	0.7	-7.0532
30	US-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Furniture	Furnishing	Howard Miller	13-3/4" Diameter Brush	124.2	3	0.2	15.525
31	US-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Envelopes	Poly String Tie	Envelopes	3.264	2	0.2	1.1016
32	US-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Art	BOSTON Model	1800 Electric Pencil Sha	86.304	6	0.2	9.7092
33	US-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Binders	Acco	Pressboard Covers with Storage Ho	6.858	6	0.7	-5.715
34	US-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Art	Lumber Crayon		15.76	2	0.2	3.546
35	CA-2017-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Appliances	Easy-staple	paper	29.472	3	0.2	9.9468
36	CA-2016-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Binders	GE 30524EE4		1097.544	7	0.2	123.4737
37	CA-2016-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Art	Electric	Architect's Clamp-On Swing Arm	190.92	5	0.6	-147.963
38	CA-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Envelopes	#10-4 1/8" x 9 1/2"	Premium Diagonal Sei	113.328	9	0.2	35.415
39	CA-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Binders	Atlantic Metals	Mobile 3-Shelf Bookcase	532.3992	3	0.32	-46.9764
40	CA-2015-1	6/9/2014		Standard	CBH-11710	Brosina Hc	Consumer	United Sta	Los Angeles	California	90032 West	Office Sup	Art	Global Fabric	Manager's Chair- Dark Gra	222.058	3	0.3	-15.147

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.

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.												
Q34 Acco Pressboard Covers with Storage Hooks- 14 7/8" x 11"- Executive Red												
G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Item Segment	Country	City	State	Postal Co	Region	Product ID	Category	Sub-Cat	Product Name	Sales	Quantity
2	re Gui Consume United Sta	Henderson Kentucky	42420 South	FUR-BO-1Furniture Bookcase	Bush Somerset Collection Bookcase	261.96	2					
3	re Gui Consume United Sta	Henderson Kentucky	42420 South	FUR-CH-1Furniture Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs- Rounded Back	731.94	3					
4	In Vai Corporate United Sta	Los Angeles California	90036 West	OFF-LA-10 Office Sup	Labels Self-Adhesive Address Labels for Typewriters by Universal	14.62	2					
5	O'DC Consume United Sta	Fort Lauderdale Florida	33311 South	FUR-TA-10 Furniture Tables	Bretford CR4500 Series Slim Rectangular Table	957.5775	5					
6	O'DC Consume United Sta	Fort Lauderdale Florida	33311 South	OFF-ST-10 Office Sup	Storage Eldon Fold 'N Roll Cart System	22.368	2					
7	Ina H Consume United Sta	Los Angeles California	90032 West	FUR-FU-1Furniture Furnishing	Eldon Expressions Wood and Plastic Desk Accessories- Cherry Wood	48.86	7					
8	Ina H Consume United Sta	Los Angeles California	90032 West	OFF-AR-1C Office Sup	Art Newell 322	7.28	4					
9	Ina H Consume United Sta	Los Angeles California	90032 West	TEC-PH-1L Technolog	Phones Mitel 5320 IP Phone VoIP phone	907.152	6					
10	Ina H Consume United Sta	Los Angeles California	90032 West	OFF-BI-10 Office Sup	Binders DXL Angle-View Binders with Locking Rings by Samsill	18.504	3					
11	Ina H Consume United Sta	Los Angeles California	90032 West	OFF-AP-1C Office Sup	Appliance Belkin F5C206VTEL 6 Outlet Surge	114.9	5					
12	Ina H Consume United Sta	Los Angeles California	90032 West	FUR-TA-10 Furniture	Tables Chromcraft Rectangular Conference Tables	1706.184	9					
13	Ina H Consume United Sta	Los Angeles California	90032 West	TEC-PH-1L Technolog	Phones Konfet 250 Conference phone - Charcoal black	911.424	4					
14	ew Al Consume United Sta	Concord North Carolina	28027 South	OFF-PA-1C Office Sup	Paper Xerox 1967	15.552	3					
15	Mad Consume United Sta	Seattle Washington	98103 West	OFF-BI-10 Office Sup	Binders Fellowes PB200 Plastic Comb Binding Machine	407.976	3					
16	ld Pa Home Offi United Sta	Fort Worth Texas	76106 Central	OFF-AP-1C Office Sup	Appliance Holmes Replacement Filter for HEPA Air Cleaner- Very Large Room- HEPA Filter	68.81	5					
17	ld 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					
18	Kriz Consume United Sta	Madison Wisconsin	53711 Central	OFF-ST-10 Office Sup	Storage Stur-D-Stor Shelving- Vertical 5-Shelf: 72" H x 36" W x 18 1/2"D	665.88	6					
19	ndro Consume United Sta	West Jord Utah	84084 West	OFF-ST-10 Office Sup	Storage Fellowes Super Stor/Drawer	55.5	2					
20	ihuuss Consume United Sta	San Francisco California	94109 West	OFF-AR-1C Office Sup	Binders Newell 341	8.56	2					
21	ihuuss Consume United Sta	San Francisco California	94109 West	TEC-PH-1L Technolog	Phones Cisco SPA 501G IP Phone	213.48	3					
22	ihuuss Consume United Sta	San Francisco California	94109 West	OFF-BI-10 Office Sup	Binders Wilson Jones Hanging View Binder- White- 1"	22.72	4					
23	Black Corporate United Sta	Fremont Nebraska	68025 Central	OFF-AP-1C Office Sup	Art Newell 318	19.46	7					
24	Black Corporate United Sta	Fremont Nebraska	68025 Central	OFF-AP-1C Office Sup	Appliance Acco Six-Outlet Power Strip- 4' Cord Length	60.34	7					
25	dra Fl Consume United Sta	Philadelphia Pennsylvania	19140 East	FUR-CH-1Furniture	Chairs Global Deluxe Stacking Chair- Gray	71.372	2					
26	y Bur Consume United Sta	Orem Utah	84057 West	FUR-TA-10 Furniture	Tables Bretford CR4500 Series Slim Rectangular Table	1044.63	3					
27	Hoffn Consume United Sta	Los Angeles California	90049 West	OFF-BI-10 Office Sup	Binders Wilson Jones Active Use Binders	11.648	2					
28	Hoffn Consume United Sta	Los Angeles California	90049 West	TEC-AC-1L Technolog	Accessori Iimation 8GB Mini TravelDrive USB 2.0 Flash Drive	90.57	3					
29	y Blur Consume United Sta	Philadelphia Pennsylvania	19140 East	FUR-BO-1Furniture	Bookcase Riverside Palais Royal Lawyers Bookcase- Royal Cherry Finish	3083.43	7					
30	y Blur Consume United Sta	Philadelphia Pennsylvania	19140 East	OFF-BI-10 Office Sup	Binders Avery Recycled Flexi-View Covers for Binding Systems	9.618	2					
31	y Blur Consume United Sta	Philadelphia Pennsylvania	19140 East	FUR-FU-1Furniture	Furnishing Howard Miller 13-3/4" Diameter Brushed Chrome Round Wall Clock	124.2	3					
32	y Blur Consume United Sta	Philadelphia Pennsylvania	19140 East	OFF-EN-1C Office Sup	Envelopes Poly String Tie Envelopes	3.264	2					
33	y Blur Consume United Sta	Philadelphia Pennsylvania	19140 East	OFF-AR-1C Office Sup	Art BOSTON Model 1800 Electric Pencil Sharpeners- Putty/Woodgrain	86.304	6					
34	y Blur Consume United Sta	Philadelphia Pennsylvania	19140 East	OFF-BI-10 Office Sup	Binders Acco Pressboard Covers with Storage Hooks- 14 7/8" x 11"- Executive Red	6.858	6					
35	y Blur Consume United Sta	Philadelphia Pennsylvania	19140 East	OFF-AP-1C Office Sup	Art Lumber Crates	15.76	2					

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.

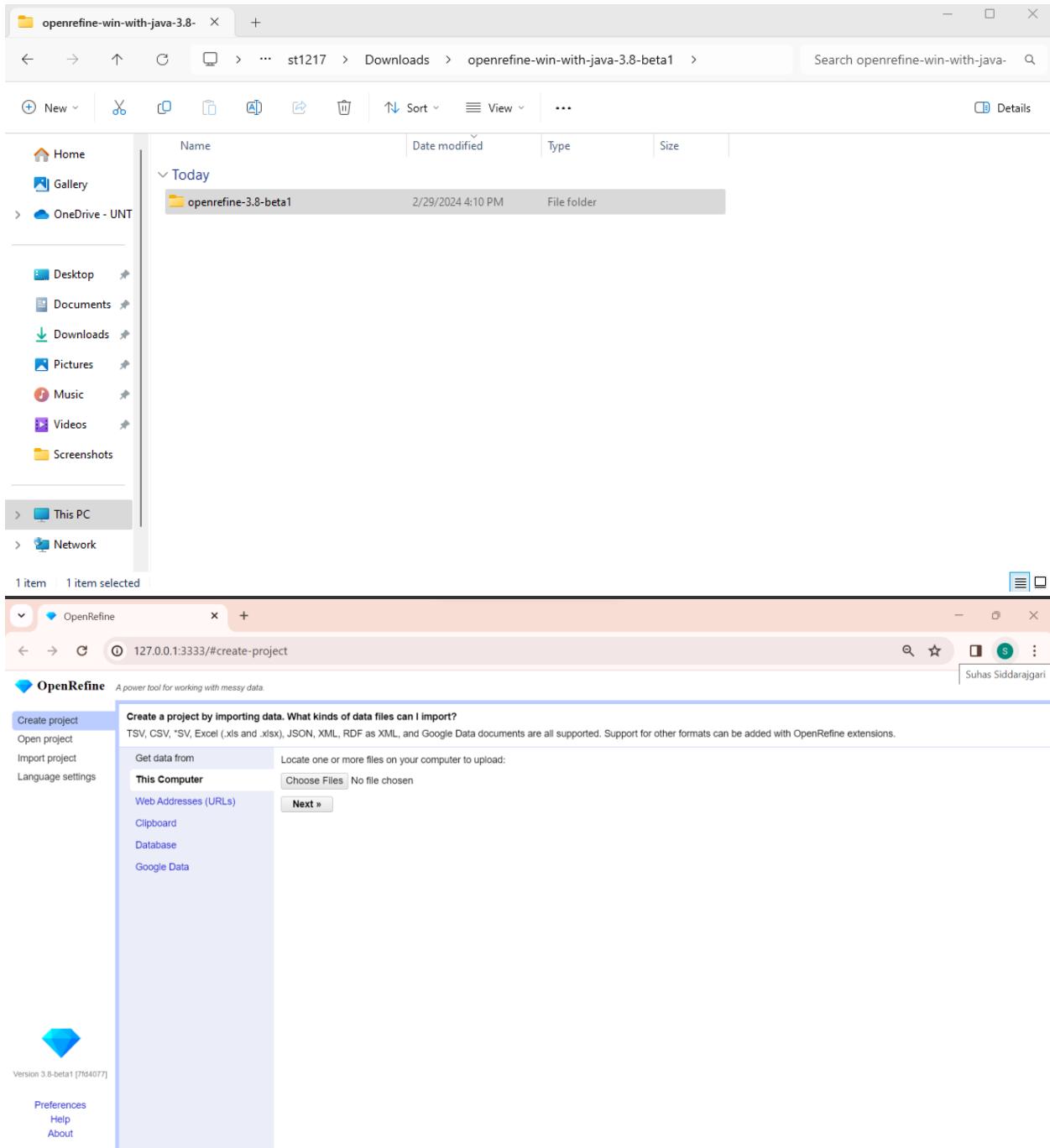
	Q	Sales	Quantity			
1	Bookcase	261.96	2			
2	Stacking Chairs- Rounded Back	731.94	3			
3	Printers for Typewriters by Universal	14.62	2			
4	Rectangular Table	957.5775	5			
5	Item	22.368	2			
6	Plastic Desk Accessories- Cherry Wood	48.86	7			
7	Phone	7.28	4			
8	Locking Rings by Samsill	907.152	6			
9	Surge	18.504	3			
10	ference Tables	114.9	5			
11	one - Charcoal black	1706.184	9			
12	mb Binding Machine	911.424	4			
13	r for HEPA Air Cleaner- Very Large Room- HEPA Filter	15.552	3			
14	Plastic Frosted Binders	407.976	3			
15	cal 5-Shelf: 72"H x 36"W x 18 1/2"D	68.81	5			
16	er	2.544	3			
17	er	665.88	6			
18	er	55.5	2			
19	er	8.56	2			
20	er	213.48	3			
21	er	22.72	4			
22	er	19.46	7			
23	er	60.34	7			
24	er	71.372	2			
25	er	1044.63	3			
26	Buri Consume United Sta Orem Utah	84057 West	FUR-TA-10 Furniture Tables	Bretford CR4500 Series Slim Rectangular Table	11.648	2
27	Hoffn Consume United Sta Los Angel California	90049 West	OFF-BI-10 Office Sup Binders	Wilson Jones Active Use Binders	90.57	3
28	Hoffn Consume United Sta Los Angel California	90049 West	TEC-AC-1 Technolog Accessori	Imation 8GB Mini TravelDrive USB 2.0 Flash Drive	3083.43	7
29	y Blur Consume United Sta Philadelphia Pennsylva	19140 East	FUR-BO-11 Furniture Bookcase	Riverside Palais Royal Lawyers Bookcase- Royale Cherry Finish	9.618	2
30	y Blur Consume United Sta Philadelphia Pennsylva	19140 East	OFF-BI-10 Office Sup Binders	Avery Recycled Flexi-View Covers for Binding Systems	124.2	3
31	y Blur Consume United Sta Philadelphia Pennsylva	19140 East	FUR-FU-11 Furniture Furnishing	Howard Miller 13-3/4" Diameter Brushed Chrome Round Wall Clock	3.264	2
32	y Blur Consume United Sta Philadelphia Pennsylva	19140 East	OFF-EN-11 Office Sup Envelopes	Poly String Tie Envelopes	86.304	6
33	y Blur Consume United Sta Philadelphia Pennsylva	19140 East	OFF-AR-11 Office Sup Art	BOSTON Model 1800 Electric Pencil Sharpeners- Putty/Woodgrain	6.858	6
34	y Blur Consume United Sta Philadelphia Pennsylva	19140 East	OFF-BI-10 Office Sup Binders	Acco Pressboard Covers with Storage Hooks- 14 7/8" x 11"- Executive Red	15.76	2
35	y Blur Consume United Sta Philadelphia Pennsylva	19140 East	OFF-AD-10 Office Sup Art	Lumber Crayons	Count: 7	100%

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

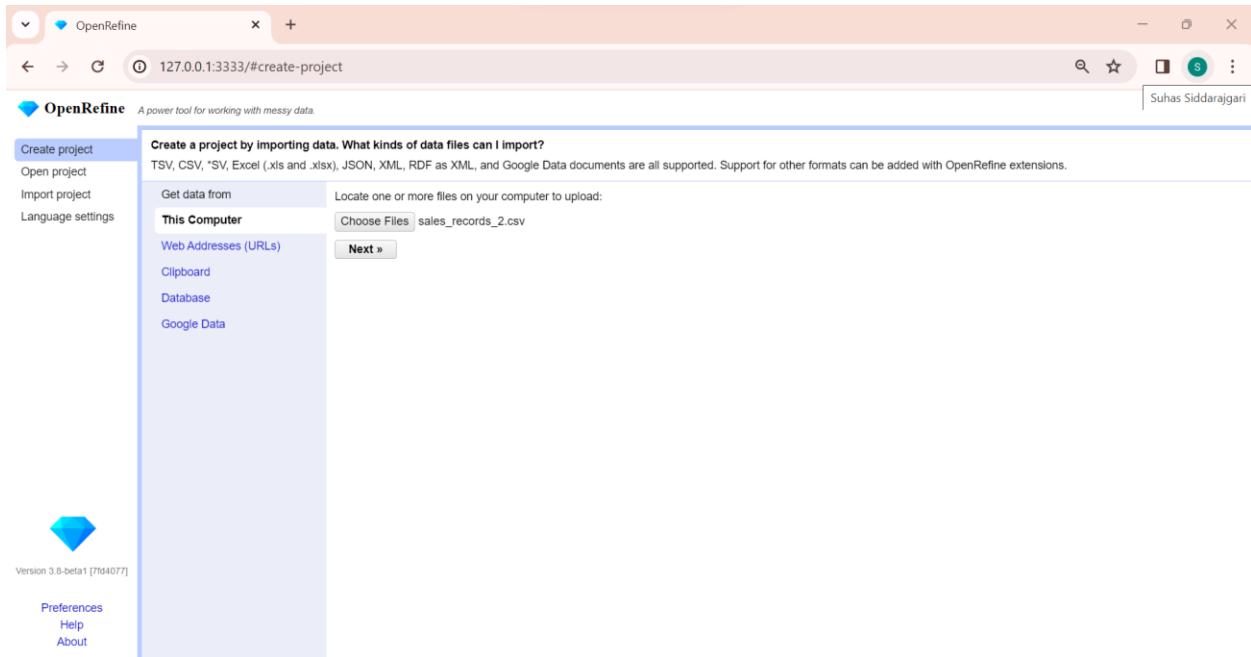
Customer Segment	Country	City	State	Postal Co	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit
Consumer	United States	Henderson	Kentucky	42420	South	FUR-BO-1	Furniture	Bookcase	Bush Somerset Collection Bookcase	261.96	2	0	41.9136
Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-1	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs- Rounded Back	731.94	3	0	219.582
Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters by Universal	14.62	2	0	6.8714
Office	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	957.5775	5	0.45	-383.031
Office	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10	Office Supplies	Storage	Eldon Fold 'N Roll Cart System	22.368	2	0.2	2.5164
Home	United States	Los Angeles	California	90032	West	FUR-FU-1	Furniture	Furnishing	Eldon Expressions Wood and Plastic Desk Accessories- Cherry Wood	48.86	7	0	14.1694
Home	United States	Los Angeles	California	90032	West	OFF-PA-1C	Office Supplies	Art	Newell 322	7.28	4	0	1.9656
Home	United States	Los Angeles	California	90032	West	TEC-PH-1	Technology	Phones	Mitel 5320 IP Phone VoIP phone	907.152	6	0.2	90.7152
Home	United States	Los Angeles	California	90032	West	OFF-BI-10	Office Supplies	Binders	DXL Angle-View Binders with Locking Rings by Samsill	18.504	3	0.2	5.7825
Home	United States	Los Angeles	California	90032	West	OFF-AP-1C	Office Supplies	Appliance	Belkin F5C206VTEL 6 Outlet Surge	114.9	5	0	34.47
Home	United States	Los Angeles	California	90032	West	FUR-TA-10	Furniture	Tables	Chromcraft Rectangular Conference Tables	1706.184	9	0.2	85.3092
Home	United States	Los Angeles	California	90032	West	TEC-PH-1	Technology	Phones	Kontrol 250 Conference phone - Charcoal black	911.424	4	0.2	68.3568
Office	United States	Concord	North Carolina	28027	South	OFF-PA-1C	Office Supplies	Paper	Xerox 1967	15.552	3	0.2	5.4432
Office	United States	Seattle	Washington	98103	West	OFF-BI-10	Office Supplies	Binders	Fellowes PB200 Plastic Comb Binding Machine	407.976	3	0.2	132.5922
Older	United States	Fort Worth	Texas	76106	Central	OFF-AP-1C	Office Supplies	Appliance	Holmes Replacement Filter for HEPA Air Cleaner- Very Large Room- HEPA Filter	68.81	5	0.8	-123.858
Older	United States	Fort Worth	Texas	76106	Central	OFF-BI-10	Office Supplies	Binders	Storex DuraTech Recycled Plastic Frosted Binders	2.544	3	0.8	-3.816
Kris	United States	Madison	Wisconsin	53711	Central	OFF-ST-10	Office Supplies	Storage	Stor-D-Stor Shelving- Vertical 5-Shelf: 72" H x 36" W x 18 1/2" D	665.88	6	0	13.3176
Indro	United States	West Jordan	Utah	84084	West	OFF-ST-10	Office Supplies	Storage	Fellowes Super Stor/Drawer	55.5	2	0	9.99
Jessus	United States	San Francisco	California	94109	West	OFF-AR-1C	Office Supplies	Art	Newell 341	8.56	2	0	2.4824
Jessus	United States	San Francisco	California	94109	West	TEC-PH-1	Technology	Phones	Cisco SPA501G IP Phone	213.48	3	0.2	16.0111
Jessus	United States	San Francisco	California	94109	West	OFF-BI-10	Office Supplies	Binders	Wilson Jones Hanging View Binder- White- 1"	22.72	4	0.2	7.384
Black	Corporate United States	Fremont	Nebraska	68025	Central	OFF-AR-1C	Office Supplies	Art	Newell 316	19.46	7	0	5.0596
Black	Corporate United States	Fremont	Nebraska	68025	Central	OFF-AP-1C	Office Supplies	Appliance	Acco Six-Outlet Power Strip- 4' Cord Length	60.34	7	0	15.6884
Dra	Consumer United States	Philadelphia	Pennsylvania	19140	East	FUR-CH-1	Furniture	Chairs	Global Deluxe Stacking Chair- Gray	71.372	2	0.3	-1.0196
Yuri	Consumer United States	Orem	Utah	84057	West	FUR-TA-10	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	1044.63	3	0	240.2649
Hoffn	Consumer United States	Los Angeles	California	90049	West	OFF-BI-10	Office Supplies	Binders	Wilson Jones Active Use Binders	11.648	2	0.2	4.2224
Hoffn	Consumer United States	Los Angeles	California	90049	West	TEC-AC-1	Technology	Accessories	Imation 8GB Mini TravelDrive USB 2.0 Flash Drive	90.57	3	0	11.7741
Yuri	Consumer United States	Philadelphia	Pennsylvania	19140	East	FUR-BO-1	Furniture	Bookcase	Riverside Palais Royal Lawyers Bookcase- Royale Cherry Finish	3083.43	7	0.5	-1665.05
Yuri	Consumer United States	Philadelphia	Pennsylvania	19140	East	OFF-BI-10	Office Supplies	Binders	Avery Recycled Flexi-View Covers for Binding Systems	9.618	2	0.7	-7.0532
Yuri	Consumer United States	Philadelphia	Pennsylvania	19140	East	FUR-FU-1	Furniture	Furnishing	Howard Miller 13-3/4" Diameter Brushed Chrome Round Wall Clock	124.2	3	0.2	15.525
Yuri	Consumer United States	Philadelphia	Pennsylvania	19140	East	OFF-EN-1	Office Supplies	Envelopes	Poly String Tie Envelopes	3.264	2	0.2	1.1016
Yuri	Consumer United States	Philadelphia	Pennsylvania	19140	East	OFF-AR-1C	Office Supplies	Art	BOSTON Model 1800 Electric Pencil Sharpeners- Putty/Woodgrain	86.304	6	0.2	9.7092
Yuri	Consumer United States	Philadelphia	Pennsylvania	19140	East	OFF-BI-10	Office Supplies	Binders	Acco Pressboard Covers with Storage Hooks- 14 7/8" x 11"- Executive Red	6.858	6	0.7	-5.715
Yuri	Consumer United States	Philadelphia	Pennsylvania	19140	East	OFF-AR-1C	Office Supplies	Art	Lumber Covers	15.76	2	0.2	3.546

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.

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Name
1.	CA-2016-152156	11/8/2016	11/11/2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-BO-10001798	Furniture	Bookcases	Bush Somers Collection Bookcase
2.	CA-2016-152156	11/8/2016	11/11/2016	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe F4 Upholstered Stacking Chair Rounded Back
3.	CA-2016-139686	6/12/2016	6/16/2016	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Label Typewriters by Universal
4.	US-2015-108966	10/11/2015	10/18/2015	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bilbao CR45 Series Slim Rectangular Ti
5.	US-2015-108966	10/11/2015	10/18/2015	Standard	SO-20335	Sean O'Donnell	Consumer	United	Fort	Florida	33311	South	OFF-ST-10000240	Office	Storage	Flinn Echt N1

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

The screenshot shows the OpenRefine interface with a sales records project. The main view displays a table with 9994 rows. A facet menu is open over the 'State' column, listing various state names like 'Alabama', 'Alaska', 'Arizona', etc., along with their frequencies. The facet menu includes options for Text facet, Numeric facet, Timeline facet, Scatterplot facet, Custom text facet, Custom numeric facet, and Customized facets.

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.

This screenshot shows the same sales records dataset in OpenRefine. The 'State' column's facet menu is open, displaying a list of unique state names from most frequent to least frequent. The menu also includes a 'Reconcile' option at the bottom.

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

sales records - OpenRefine

127.0.0.1:3333/project?project=2017061710567

Suhas Siddarajgarj | Open... Export Help

OpenRefine sales records Permalink

Facet / Filter Undo / Redo 0 / 0

9994 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

State Ship Mode Customer ID Customer Name Segment Country City State Postal Code Region Product ID Category Sub-Category Pn

State	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Pn
Alabama 61	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-B0-10001798	Furniture	Bookcases	Bush S Collect Bookc
Arizona 224	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon D Uphols Stackr Round
Arkansas 60	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Adr Address Typew Univer
California 2001	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bretfor Series Rectar
Colorado 182	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon I Cart St
Connecticut 82	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon I Wood : Desk A Cherry
Delaware 96	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Technology	Mitel 5 Phone
District of Columbia 10	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Office Supplies	Phones	DXL Ar Binden Lockin Samsl
Florida 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	Belkin F5C20 Outlet Chrom Duetar
Georgia 184	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Newell
Idaho 21	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10000240	Furniture	Tables	
Illinois 492	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				
Indiana 149	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				
Iowa 30	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				
Kansas 24	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				
Kentucky 139	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				
Louisiana 42	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				
Maine 8	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				
Maryland 105	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				
Massachusetts 135	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				
Michigan 255	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				
Minnesota 89	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				
Mississippi 53	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				

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.

sales records - OpenRefine

127.0.0.1:3333/project?project=2017061710567

Suhas Siddarajgarj | Open... Export Help

OpenRefine sales records Permalink

Facet / Filter Undo / Redo 0 / 0

9994 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

State Ship Mode Customer ID Customer Name Segment Country City State Postal Code Region Product ID Category Sub-Category Pn

State	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Pn
North Dakota 7	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-B0-10001798	Furniture	Bookcases	Bush S Collect Bookc
West Virginia 4	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon D Uphols Stackr Round
Floreda 1	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Adr Address Typew Univer
Nebreska 1	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bretfor Series Rectar
Texs 1	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon I Cart St
Wyoming 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon I Wood : Desk A Cherry
Facet by choice counts	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Technology	Mitel 5 Phone
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Office Supplies	Phones	DXL Ar Binden Lockin Samsl
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10000240	Furniture	Tables	

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

State	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Prod
Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-B0-10001798	Furniture	Bookcases	Bush S Collect Bookc	
Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon D Uphols Stackr Round	
Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Ad Address Typewi Univer	
Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bretfor Series Rectar	
Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon I Cart S	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon I Wood : Desk A Cherry Newell	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	Mitel 5 Phone	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Technology	Phones	DXL Ar Binden Lockin Samsl	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	Belkin F5C20 Outlet	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Chrom Dauter	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10000530	Furniture	Tables		

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

State	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Prod
Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-B0-10001798	Furniture	Bookcases	Bush S Collect Bookc	
Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon D Uphols Stackr Round	
Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Ad Address Typewi Univer	
Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bretfor Series Rectar	
Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon I Cart S	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon I Wood : Desk A Cherry Newell	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	Mitel 5 Phone	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Technology	Phones	DXL Ar Binden Lockin Samsl	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	Belkin F5C20 Outlet	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Chrom Dauter	
Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10000530	Furniture	Tables		

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

State	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Pn
Maine 8	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-BO-10001798	Furniture	Bookcases	Bush S Collect Bookc
North Dakota 7	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon D Uphols Stackir Round
West Virginia 4	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Ad Address Typewi Univer
Nebraska 1	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bretfor Series Rectar
Texas 1	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon I Cart S
Wyoming 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon I Wood : Desk A Cherry Newell
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Technology	Phones	Mitel 5 Phone
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	DXL Ar Bindr Lockin Samll
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C20 Outlet
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10000520	Furniture	Tables	Chrom Chrom

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

State	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Pn
Maine 8	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-BO-10001798	Furniture	Bookcases	Bush S Collect Bookc
North Dakota 7	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon D Uphols Stackir Round
West Virginia 4	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Ad Address Typewi Univer
Nebraska 1	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bretfor Series Rectar
Texas 1	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon I Cart S
Wyoming 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon I Wood : Desk A Cherry Newell
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Technology	Phones	Mitel 5 Phone
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	DXL Ar Bindr Lockin Samll
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C20 Outlet
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10000520	Furniture	Tables	Chrom Chrom

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

The screenshot shows the OpenRefine interface with a project titled "sales records". The facet for "State" is open, showing a list of states with their counts: District of Columbia (10), Maine (8), North Dakota (7), West Virginia (4), Texas (1), and Wyoming (1). The entry "Texas" is highlighted. The main data table shows 9994 rows of sales records. The "Postal Code" column is currently sorted by value.

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

The screenshot shows the OpenRefine interface with the same project. The context menu for the "Postal Code" column is open, and the "Numeric facet" option is selected. The main data table remains the same, showing 9994 rows of sales records.

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.

Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Pn
	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-BO-10001798	Furniture	Bookcases	Bush S Collect Bookc
	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon D Uphols Stackir Round
	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Ad Address Typew Univer
	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Bretfor Series Rectar
	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon I Cart S
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon I Wood : Desk A Cherry Newell
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	Mitel 5 Phone
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Technology	Phones	DXL Ar Binden Lockin Samsl
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	Belkin F5C20 Outlet Chrom
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C20 Outlet Chrom
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10000570	Furniture	Tables	Chrom

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

Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Pn
	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	Facet					Bookcases
	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	Text filter					Chairs
	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	33311					Labels
	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South				Tables
	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South				Storage
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				Furnishings
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				Art
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				Phones
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				Binders
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				Appliances
	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West				Belkin F5C20 Outlet Chrom

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

Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Pn
10009 229	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-B0-10001798	Furniture	Bookcases	Bush S Collect Bookcas
10011 193	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon D Uphols Stackir Round
10024 230	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Adr Address Typewr Univer
10035 263	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Brefor Series Rectar
1040 1	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon I Cart St
10550 8	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon I Wood Desk A Cherry
10701 15	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	Newell
10801 9	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Technology	Phones	Mitel 5 Phone
11520 6	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	DXL Ar Binden Lockin Samsil
11550 11	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C20 Outlet
11561 34	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10000240	Furniture	Tables	Chrom D

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.

Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Pn
83501 1	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-B0-10001798	Furniture	Bookcases	Bush S Collect Bookcas
8401 1	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	Kentucky	42420	South	FUR-CH-10000454	Furniture	Chairs	Hon D Uphols Stackir Round
84041 1	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	West	OFF-LA-10000240	Office Supplies	Labels	Self-Adr Address Typewr Univer
90604 1	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	FUR-TA-10000577	Furniture	Tables	Brefor Series Rectar
90640 1	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South	OFF-ST-10000760	Office Supplies	Storage	Eldon I Cart St
91761 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-FU-10001487	Furniture	Furnishings	Eldon I Wood Desk A Cherry
92253 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	Newell
92399 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Technology	Phones	Mitel 5 Phone
92530 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	DXL Ar Binden Lockin Samsil
93405 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C20 Outlet
93454 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10000240	Furniture	Tables	Chrom D
94061 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	Newell
94403 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Technology	Phones	Mitel 5 Phone
94509 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	DXL Ar Binden Lockin Samsil
94568 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C20 Outlet
95610 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10000240	Furniture	Tables	Chrom D
95687 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	Newell
96003 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	TEC-PH-10002275	Technology	Phones	Mitel 5 Phone
98002 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-BI-10003910	Office Supplies	Binders	DXL Ar Binden Lockin Samsil
98208 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AP-10002892	Office Supplies	Appliances	Belkin F5C20 Outlet
central 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	FUR-TA-10000240	Furniture	Tables	Chrom D
Pennsylvania 1	Standard Class	BH-11710	Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	West	OFF-AR-10002833	Office Supplies	Art	Newell

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.

The screenshot shows the OpenRefine interface with a project titled "sales records". A context menu is open over a row of data, specifically for the "Postal Code" column. The menu path "Edit cells > Common transforms > To number" is highlighted. The data table contains 9994 rows and various columns like "Customer ID", "Customer Name", "Segment", etc. The "Postal Code" column initially contains text values like "FUR-BO-10001798".

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.

This screenshot shows the OpenRefine interface after the conversion. The "Postal Code" facet panel is visible on the left, featuring a histogram with a scale from 1,000 to 100,000. A checkbox labeled "Numeric" is checked, indicating that the data has been successfully converted to a numeric type. The main data table now displays numeric postal codes in the "Postal Code" column, such as 42420, 90036, etc.

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.

Date	Ship Mode	Customer ID	Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Prod
Standard Class	DR-12880	Dan Reichenbach	Corporate	United States	Chicago	Illinois	central	OFF-ST-10004804	Central	Office Supplies	Storage	Belkin 19" Equipment Shelf-Black	
Standard Class	AG-10495	Andrew Gjertsen	Corporate	United States	Philadelphia	Pennsylvania	Pennsylvania	OFF-ST-10000876	East	Office Supplies	Storage	Eldon Simplefile Box Office	

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

Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit
Corporate	United States	Chicago	Illinois	central	edit	OFF-ST-10004804	Office Supplies	Storage	Belkin 19" Vented Equipment Shelf-Black	82.368	2	0.2	-19.5624
Corporate	United States	Philadelphia	Pennsylvania	Pennsylvania	Edit this cell	OFF-ST-10000876	Office Supplies	Storage	Eldon Simplefile Box Office	59.712	6	0.2	5.9712

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.

The screenshot shows the OpenRefine interface with a project titled "sales records". The "Postal Code" facet is selected, showing a histogram of postal codes between 1,000 and 100,000. A modal dialog is open over a row for Andrew Gjertsen, where the postal code "19140" is being converted from "Non-numeric" to "number". The row details are:

Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity
Dan Reichenbach	Corporate	United States	Chicago	Illinois	60610	Central	OFF-ST-10004804	Office Supplies	Storage	Belkin 19" Vented Equipment Shelf-Black	82.368	2
Andrew Gjertsen	Corporate	United States	Philadelphia	Pennsylvania	Pennsylvania						59.712	6

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

The screenshot shows the OpenRefine interface with the same project and data. The "Postal Code" facet is still selected. The row for Andrew Gjertsen now shows the corrected postal code "19140" in the "Postal Code" column. The row details are:

Customer Name	Segment	Country	City	State	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit
Dan Reichenbach	Corporate	United States	Chicago	Illinois	60610	Central	OFF-ST-10004804	Office Supplies	Storage	Belkin 19" Vented Equipment Shelf-Black	82.368	2	0.2	-19.5624
Andrew Gjertsen	Corporate	United States	Philadelphia	Pennsylvania	19140	East	OFF-ST-10000876	Office Supplies	Storage	Eldon Simplefile Box Office	59.712	6	0.2	5.9712

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.

Customer Name	Segment	Country	City	State	Postal Code	Postal Code 2	Region	Product ID	Category	Sub-Category	Product Name	
Sally Knutson	Consumer	United States	Fairfield	Connecticut	6824	99999	East	OFF-BI-10003470	Office Supplies	Binders	Avery Poly Binder Pockets	7.1
Valerie Mitchell	Home Office	United States	Westfield	New Jersey	7090	99999	East	OFF-ST-10001414	Office Supplies	Storage	Decofiles Hanging Personal Folder File	46
Paul Gonzalez	Consumer	United States	Morristown	New Jersey	7960	99999	East	OFF-FA-10003472	Office Supplies	Fasteners	Bagged Rubber Bands	7.5
Jonathan Doherty	Corporate	United States	Belleville	New Jersey	7109	99999	East	OFF-PA-10002105	Office Supplies	Paper	Xerox 223	32
Jonathan Doherty	Corporate	United States	Belleville	New Jersey	7109	99999	East	OFF-ST-10002756	Office Supplies	Storage	Tennsco Stur-D-Stor Boltless Shelving- 5 Shelves 24" Deep-Sand	10
Jonathan Doherty	Corporate	United States	Belleville	New Jersey	7109	99999	East	OFF-PA-10004243	Office Supplies	Paper	Xerox 1939	56
Jonathan Doherty	Corporate	United States	Belleville	New Jersey	7109	99999	East	FUR-FU-10001861	Furniture	Furnishings	Floodlight Indoor Halogen Bulbs- 1 Bulb per Pack- 60 Watts	77
Jonathan Doherty	Corporate	United States	Belleville	New Jersey	7109	99999	East	OFF-BI-10002706	Office Supplies	Binders	Avery Premier Heavy-Duty Binder with Round Locking Rings	14
Corey Roper	Home Office	United States	Lakewood	New Jersey	8701	99999	East	OFF-BI-10001072	Office Supplies	Binders	GBC Clear Cover- 8-1/2 x 11- unprinted- 25 covers per pack	45
Corey Roper	Home Office	United States	Lakewood	New Jersey	8701	99999	East	OFF-AR-10002135	Office Supplies	Art	Boston Heavy-Duty Trimline Electric Pencil Sharpeners	28
Thomas Seio	Corporate	United States	Hackensack	New Jersey	7601	99999	East	FUR-FU-10001860	Furniture	Furnishings	Deflect-O CounterTop	87

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.

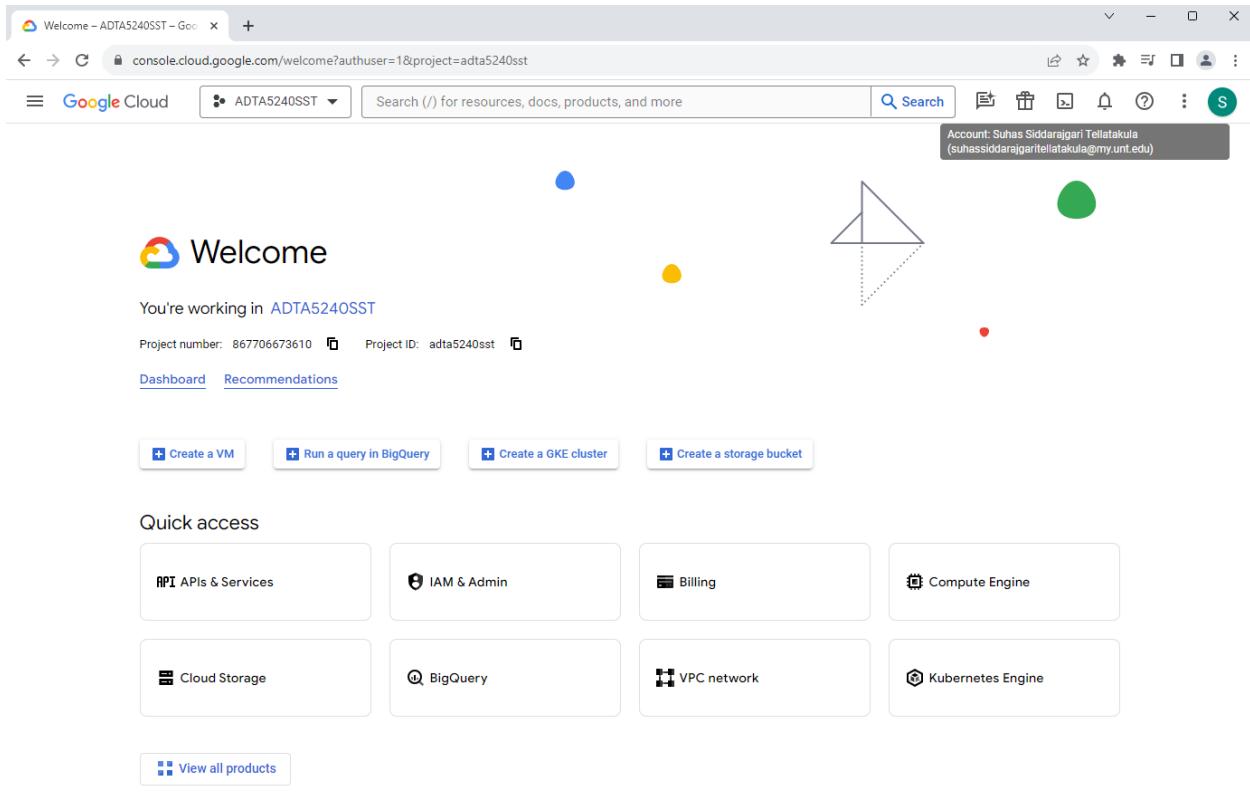
The screenshot shows the OpenRefine interface with a table titled "sales records - OpenRefine". The table contains 9994 rows of data with columns: Customer Name, Segment, Country, City, State, Postal Code, Postal Code 2, Region, Product ID, and Category. The Category column shows various product categories like Furniture, Office Supplies, Technology, and Appliances. The export menu is open on the right, with "Comma-separated value" selected. Other options include "OpenRefine project archive to file", "Tab-separated value", "HTML table", "Excel (.xls)", "Excel 2007+ (.xlsx)", "ODF spreadsheet", "Custom tabular...", "SQL...", "Templating...", "OpenRefine project archive to Google Drive...", "Google Sheets...", "Wikibase edits...", "QuickStatements file", and "Wikibase schema".

Customer Name	Segment	Country	City	State	Postal Code	Postal Code 2	Region	Product ID	Category
Claire Gute	Consumer	United States	Henderson	Kentucky	42420	42420	South	FUR-BO-10001798	Furniture
Claire Gute	Consumer	United States	Henderson	Kentucky	42420	42420	South	FUR-CH-10000454	Furniture
Darrin Van Huff	Corporate	United States	Los Angeles	California	90036	90036	West	OFF-LA-10000240	Office Supplies
Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	33311	South	FUR-TA-10000577	Furniture
Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	33311	South	OFF-ST-10000760	Office Supplies
Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	90032	West	FUR-FU-10001487	Furniture
Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	90032	West	OFF-AR-10002833	Office Supplies
Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	90032	West	TEC-PH-10002275	Technology
Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	90032	West	OFF-BI-10003910	Office Supplies
Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	90032	West	OFF-AP-10002892	Appliances
Brosina Hoffman	Consumer	United States	Los Angeles	California	90032	90032	West	FUR-TA-10001420	Furniture

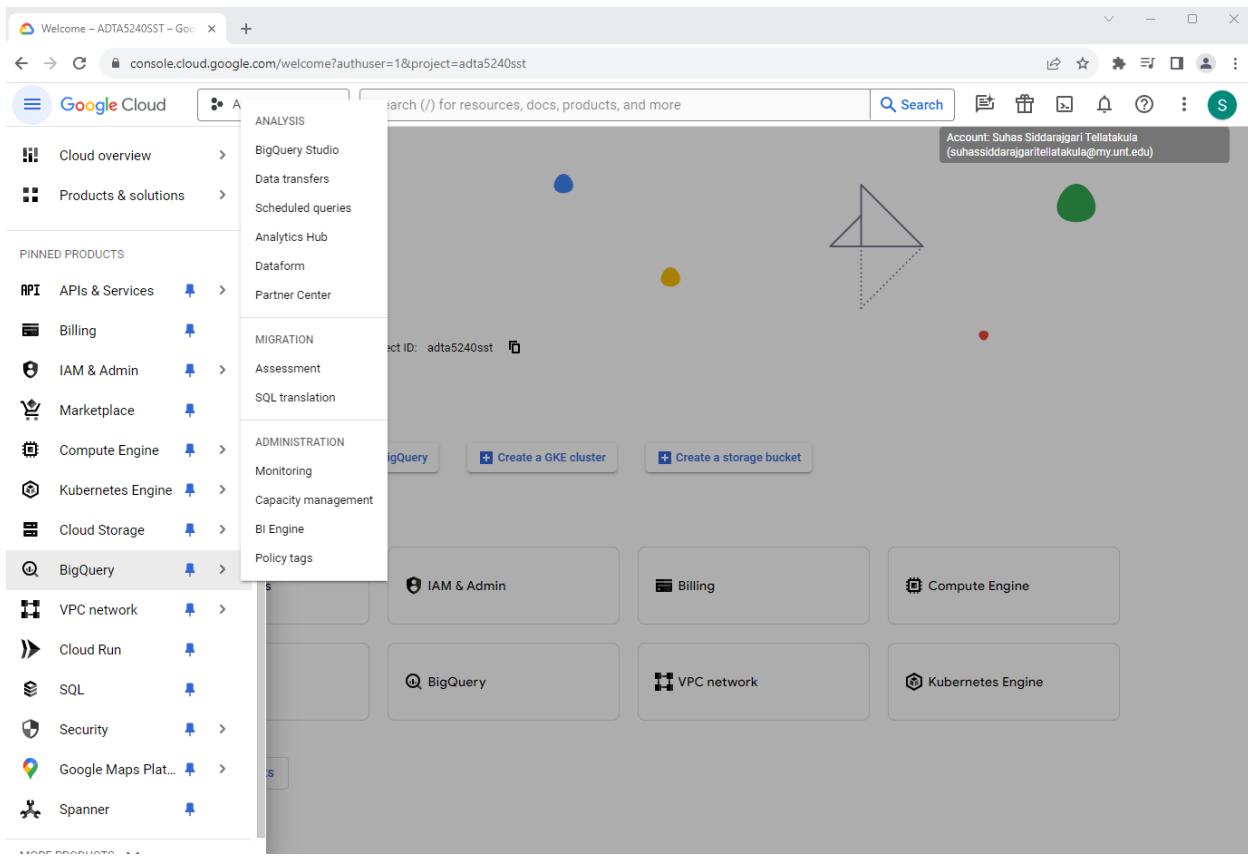
Part 4 – Data Analysis

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.

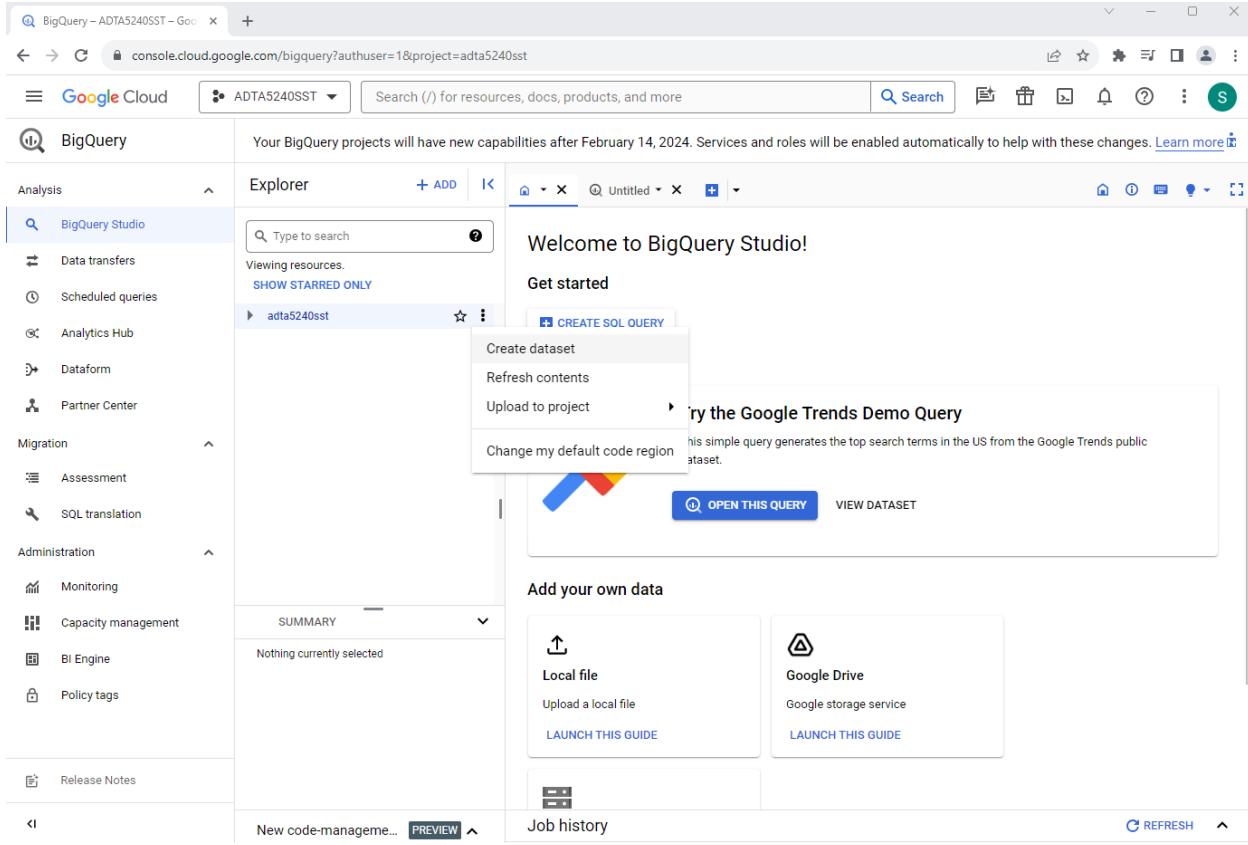
First, I signed into my GCP console which landed me in project home page as shown in below screenshot.



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.

The screenshot shows the Google Cloud BigQuery 'Create dataset' dialog box. On the left, the main interface displays the 'BigQuery Studio' sidebar with options like Analysis, Data transfers, Scheduled queries, Analytics Hub, Dataform, Partner Center, Migration, Assessment, SQL translation, Administration, Monitoring, Capacity management, BI Engine, Policy tags, and Release Notes. The 'Job history' section at the bottom right is partially visible. The central area shows a 'Welcome to BigQuery' card with a 'CREATE SQL QUERY' button and a 'Try the sample data' section. The 'Create dataset' dialog is open on the right, containing the following fields:

- Project ID:** adta5240sst (with a 'CHANGE' link)
- Dataset ID ***: salesrecords
- Location type:** Multi-region (Allow BigQuery to select a region within a group to achieve higher quota limits.)
- Multi-region *:** US (multiple regions in United States)
- Default table expiration:** Enable table expiration (Days field)
- Advanced options:** A dropdown menu is open here.

At the bottom of the dialog are 'CREATE DATASET' and 'CANCEL' buttons.

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.

The screenshot shows the Google Cloud BigQuery Dataset Info page for the 'salesrecords' dataset. The dataset was created on Feb 29, 2024, at 6:43:41 PM UTC-6. It is located in the US and has a default rounding mode of ROUNDING_MODE_UNSPECIFIED. The dataset replica info shows a primary location in the US with 1 replica. A success message 'salesrecords* created.' is displayed at the bottom of the page.

Dataset ID	adta5240sst.salesrecords
Created	Feb 29, 2024, 6:43:41 PM UTC-6
Default table expiration	Never
Last modified	Feb 29, 2024, 6:43:41 PM UTC-6
Data location	US
Description	
Default collation	
Default rounding mode	ROUNDING_MODE_UNSPECIFIED
Time travel window	7 days
Storage billing model	LOGICAL
Case insensitive	false
Labels	
Tags	

Dataset replica info [PREVIEW](#) [VIEW REPLICAS](#)

Primary location	US
Replicas	1

"salesrecords* created. [GO TO DATASET](#) X

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!m2!1sadta5240sst!2ssalesrecords

Create table

Source

- Create table from
- Upload

Select file *
Cleaned_records_3.csv

File format
CSV

Destination

Project * adta5240sst

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 - BigQuery - ADTA

console.cloud.google.com/bigquery?authuser=1&project=adta5240sst&ws=!1m4!1m3!m2!1sadta5240sst!2ssalesrecords

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](#)

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

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.

The screenshot shows the Google Cloud BigQuery interface. On the left, the sidebar includes sections for Analysis, Data transfers, Scheduled queries, Analytics Hub, Dataform, Partner Center, Migration, Assessment, SQL translation, Administration, Monitoring, Capacity management, BI Engine, and Policy tags. A 'Release Notes' section is also present. The main area is titled 'Explorer' and shows a tree view of datasets and tables. Under the 'salesrecords' dataset, the 'salesanalysis' table is selected. A modal window at the bottom center displays the schema of the 'salesanalysis' table:

Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
Row_ID	INTEGER	NULLABLE	-	-	-	-	-
Order_ID	STRING	NULLABLE	-	-	-	-	-
Order_Date	DATE	NULLABLE	-	-	-	-	-
Ship_Date	DATE	NULLABLE	-	-	-	-	-
Ship_Mode	STRING	NULLABLE	-	-	-	-	-
Customer_ID	STRING	NULLABLE	-	-	-	-	-
Customer_Name	STRING	NULLABLE	-	-	-	-	-
Segment	STRING	NULLABLE	-	-	-	-	-
Country	STRING	NULLABLE	-	-	-	-	-
City	STRING	NULLABLE	-	-	-	-	-
State	STRING	NULLABLE	-	-	-	-	-
Postal_Code	INTEGER	NULLABLE	-	-	-	-	-
Postal_Code_2	INTEGER	NULLABLE	-	-	-	-	-
Region	STRING	NULLABLE	-	-	-	-	-
Product_ID	STRING	NULLABLE	-	-	-	-	-
Category	STRING	NULLABLE	-	-	-	-	-
Sub_Category	STRING	NULLABLE	-	-	-	-	-
Product_Name	STRING	NULLABLE	-	-	-	-	-
Sales	FLOAT	NULLABLE	-	-	-	-	-
Quantity	INTEGER	NULLABLE	-	-	-	-	-
Discount	FLOAT	NULLABLE	-	-	-	-	-
Profit	FLOAT	NULLABLE	-	-	-	-	-

Below the schema, there is a 'SUMMARY' section with details about the table:

- Last modified: Feb 29, 2024, 7:14:46 PM UTC-6
- Data location: US
- Description:
- Labels:
- Table type: table

A modal window at the bottom center says "salesanalysis* created." with buttons for "GO TO TABLE" and "X".

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

The screenshot shows the Google Cloud BigQuery console interface. On the left, there's a sidebar with sections like Analysis, Migration, and Administration. The main area is titled 'salesanalysis' and shows a preview of the data. The schema includes columns for Customer_ID, Customer_Name, Segment, Country, and City. The preview table contains 15 rows of sample data. At the bottom, there are navigation controls for results per page (50), a refresh button, and a job history link.

Row	Customer_ID	Customer_Name	Segment	Country	City
1	NW-18400	Natalie Webber	Consumer	United States	Waterloo
2	CS-12175	Charles Sheldon	Corporate	United States	Iowa City
3	DE-13255	Deanna Eno	Home Office	United States	Dubuque
4	DE-13255	Deanna Eno	Home Office	United States	Dubuque
5	RD-19585	Rob Dowd	Consumer	United States	Dubuque
6	ML-17755	Max Ludwig	Home Office	United States	Marion
7	PW-19030	Pauline Webber	Corporate	United States	Marion
8	TM-21010	Tamara Manning	Consumer	United States	Marion
9	TM-21010	Tamara Manning	Consumer	United States	Marion
10	TM-21010	Tamara Manning	Consumer	United States	Marion
11	TM-21010	Tamara Manning	Consumer	United States	Marion
12	PO-18850	Patrick O'Reilly	Consumer	United States	Burlington
13	RS-19870	Roy Skaria	Home Office	United States	Burlington
14	RS-19870	Roy Skaria	Home Office	United States	Burlington
15	BV-11245	Benjamin Venier	Corporate	United States	Des Moines

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.

Row	Customer_ID	Customer_Name	Segment	Country	City
1	NW-18400	Roxanne Webber	Consumer	United States	Waterloo
2	CS-12175	Charles Sheldon	Corporate	United States	Iowa City
3	DE-13255	Deanna Eno	Home Office	United States	Dubuque
4	DE-13255	Deanna Eno	Home Office	United States	Dubuque
5	RD-19585	Rob Dowd	Consumer	United States	Dubuque
6	ML-17755	Max Ludwig	Home Office	United States	Marion
7	PW-19030	Pauline Webber	Corporate	United States	Marion
8	TM-21010	Tamara Manning	Consumer	United States	Marion
9	TM-21010	Tamara Manning	Consumer	United States	Marion
10	TM-21010	Tamara Manning	Consumer	United States	Marion
11	TM-21010	Tamara Manning	Consumer	United States	Marion
12	PO-18850	Patrick O'Reilly	Consumer	United States	Burlington
13	RS-19870	Roy Skaria	Home Office	United States	Burlington
14	RS-19870	Roy Skaria	Home Office	United States	Burlington
15	BV-11245	Benjamin Venier	Corporate	United States	Des Moines

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.

The screenshot shows the Google Cloud BigQuery interface. On the left, the sidebar includes sections for Analysis, Migration, and Assessment. The main area is titled 'BigQuery Studio' under 'Analysis'. It features an 'Explorer' pane with a search bar and a tree view showing a project named 'adta5240sst' containing a dataset 'salesrecords' and a table 'salesanalysis'. To the right, a query editor window is open with the title 'Untitled 2'. The query code is:

```

1 SELECT State, COUNT(DISTINCT Customer_ID) AS NumCustomers
2 FROM salesrecords.salesanalysis
3 GROUP BY State
4 ORDER BY NumCustomers DESC
5 LIMIT 5;

```

The status bar at the bottom of the query editor says 'Query completed.' Below the query editor is a 'Query results' table with the following data:

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

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

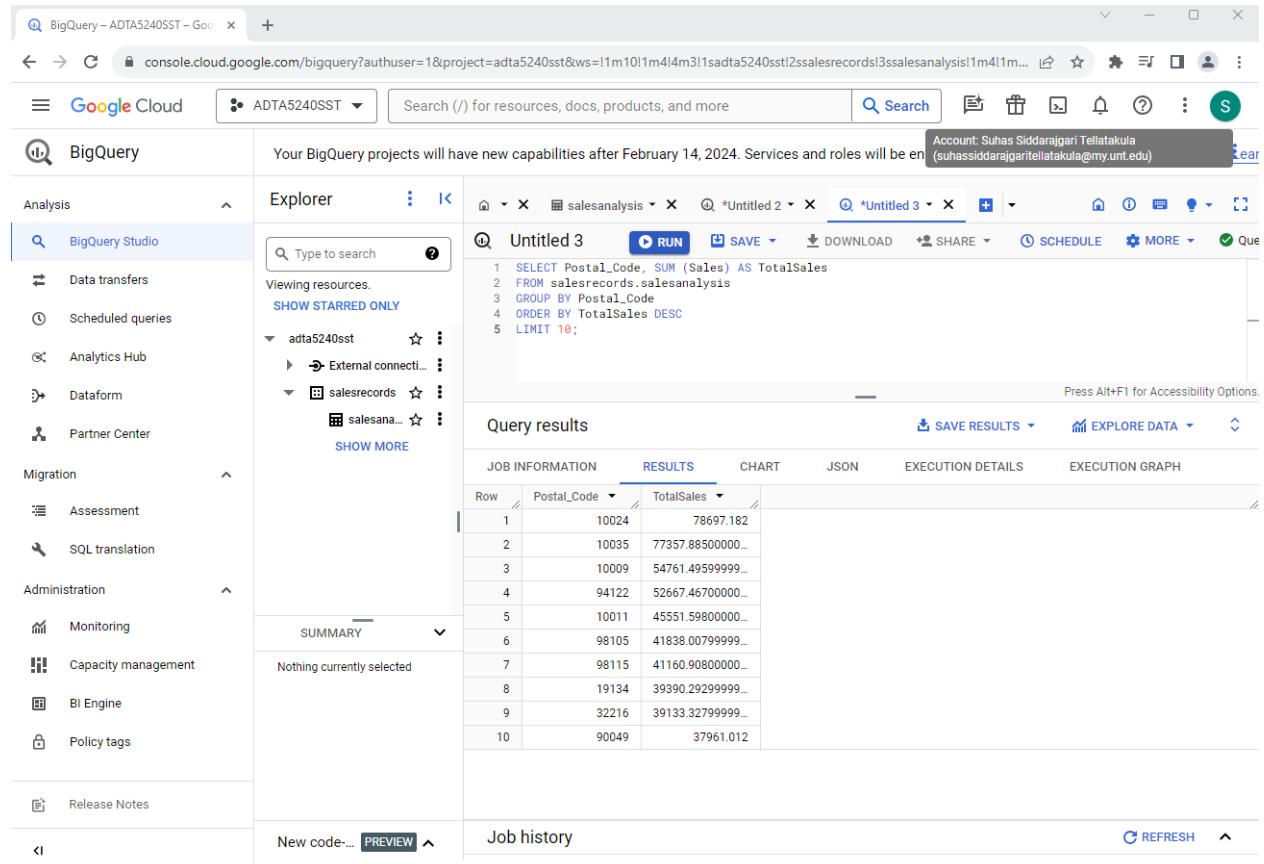
FROM salesrecords.salesanalysis

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.



The screenshot shows the Google Cloud BigQuery interface. On the left, there's a sidebar with sections like Analysis, Migration, and Administration. The main area has tabs for Explorer, Query results, and Job history. A query titled 'Untitled 3' is running, displaying the following SQL code:

```
1 SELECT Postal_Code, SUM (Sales) AS TotalSales
2 FROM salesrecords.salesanalysis
3 GROUP BY Postal_Code
4 ORDER BY TotalSales DESC
5 LIMIT 10;
```

The 'Query results' tab is selected, showing a table with two columns: 'Postal_Code' and 'TotalSales'. The data is as follows:

Row	Postal_Code	TotalSales
1	10024	78697.182
2	10035	77357.88500000...
3	10009	54761.49599999...
4	94122	52667.46700000...
5	10011	45551.59800000...
6	98105	41838.00799999...
7	98115	41160.90800000...
8	19134	39390.29299999...
9	32216	39133.32799999...
10	90049	37961.012

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	54761.49599999963
4	94122	52667.467000000011
5	10011	45551.598000000013
6	98105	41838.0079999998
7	98115	41160.90800000001
8	19134	39390.29299999976
9	32216	39133.32799999994
10	90049	37961.012