Advanced Big Data: Final Exam - Project

Name: **SARAVANAN SOUNDARYA SUNDARI**

ID: **2021120203**

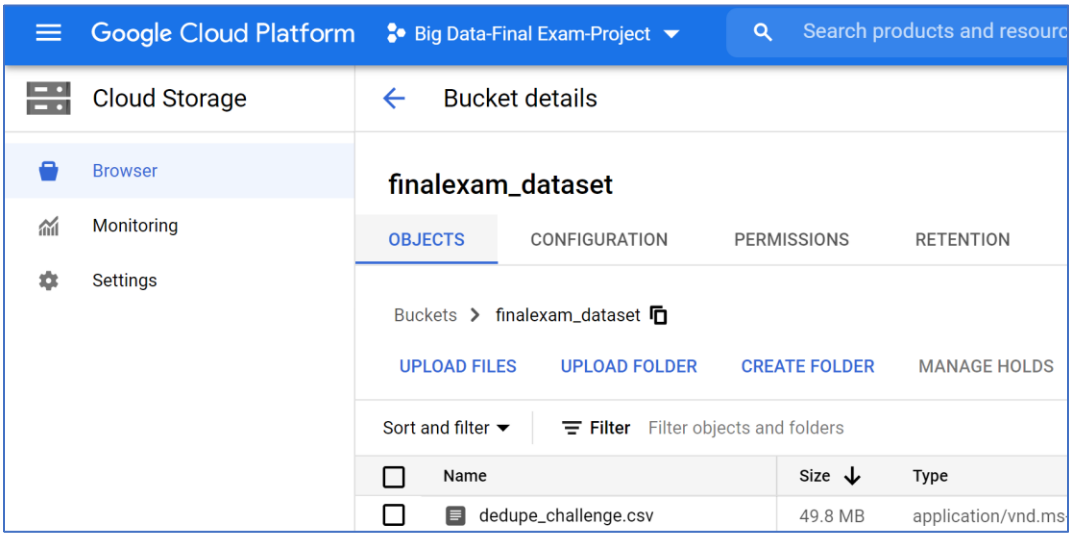
Google Cloud Platform (GCP)

Setting up Google Cloud Platform (GCP) to perform Data Cleaning.

Stage Data in Cloud Storage

In **Cloud Storage**, create bucket ‘finalexam_dataset’.

Stage data ‘dedupe_challenge.csv’ in ‘finalexam_dataset’ bucket, in Cloud Storage. (refer Fig.1)



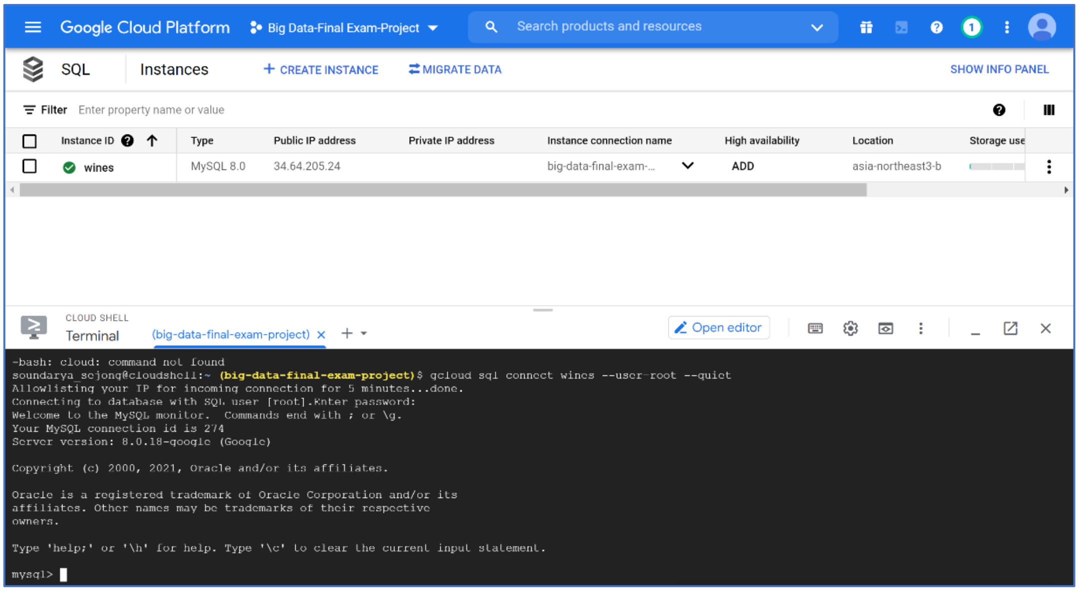
□	Name	Size ↓	Type
□	dedupe_challenge.csv	49.8 MB	application/vnd.ms

Fig.1

Create Database in Cloud SQL

In **Cloud SQL**, created SQL instance with Instance Id as ‘wines’. Connected to ‘wines’ instance using **Cloud Shell.** to connect to Database. (refer Fig 2.)

```
gcloud sql connect wines --user=root --quiet
```

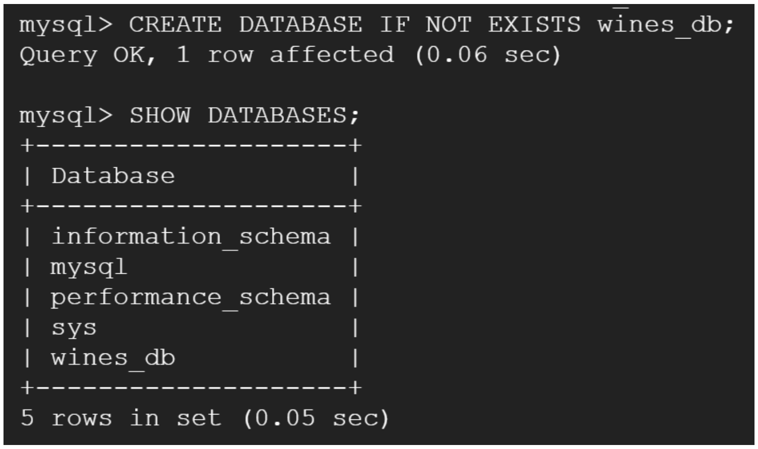

Google Cloud Platform
Big Data-Final Exam-Project
Search products and resources
SQL
Instances
CREATE INSTANCE
MIGRATE DATA
SHOW INFO PANEL
Filter
Enter property name or value
CLOUD SHELL
Terminal
(big-data-final-exam-project)
Open editor

Instance ID	Type	Public IP address	Private IP address	Instance connection name	High availability	Location	Storage use
wines	MySQL 8.0	34.64.205.24		big-data-final-exam-...	ADD	asia-northeast3-b	

```
-bash: cloud: command not found
soundarya_scjong@cloudshell:~ (big-data-final-exam-project)$ gcloud sql connect wines --user=root --quiet
Allowlisting your IP for incoming connection for 5 minutes...done.
Connecting to database with SQL user [root].Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 274
Server version: 8.0.18-google (Google)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> 
```

Fig. 2

Create Database ‘wines_db’ in Cloud Shell. (refer Fig 3.)



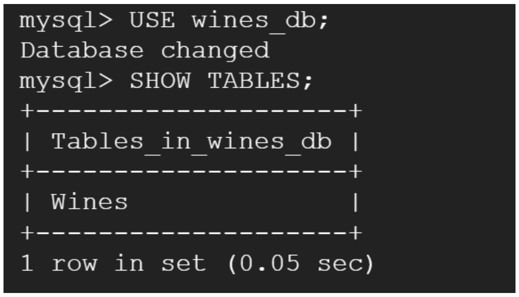
```
mysql> CREATE DATABASE IF NOT EXISTS wines_db;
Query OK, 1 row affected (0.06 sec)

mysql> SHOW DATABASES;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mysql              |
| performance_schema |
| sys                |
| wines_db           |
+--------------------+
5 rows in set (0.05 sec)
```

Fig. 3

Create table ‘Wines’. (refer Fig. 4, Fig. 5)

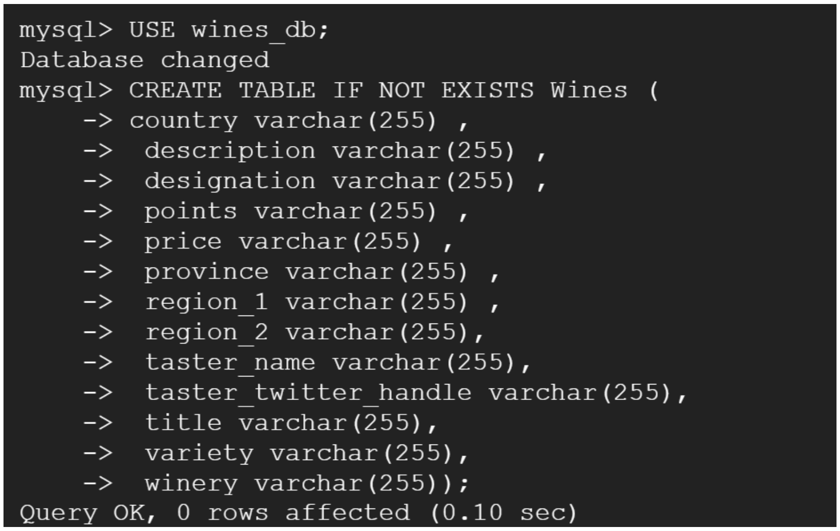
SQL Code:

```
CREATE TABLE IF NOT EXISTS Accommodation
(country varchar(255) ,
 description varchar(255) ,
designation varchar(255) ,
points varchar(255) ,
price varchar(255) ,
province varchar(255) ,
region_1 varchar(255) ,
region_2 varchar(255),
taster_name varchar(255),
taster_twitter_handle varchar(255),
title varchar(255),
variety varchar(255),
winery varchar(255));
```



```
mysql> USE wines_db;
Database changed
mysql> SHOW TABLES;
+--------------------+
| Tables_in_wines_db |
+--------------------+
| Wines              |
+--------------------+
1 row in set (0.05 sec)
```

Fig 4.

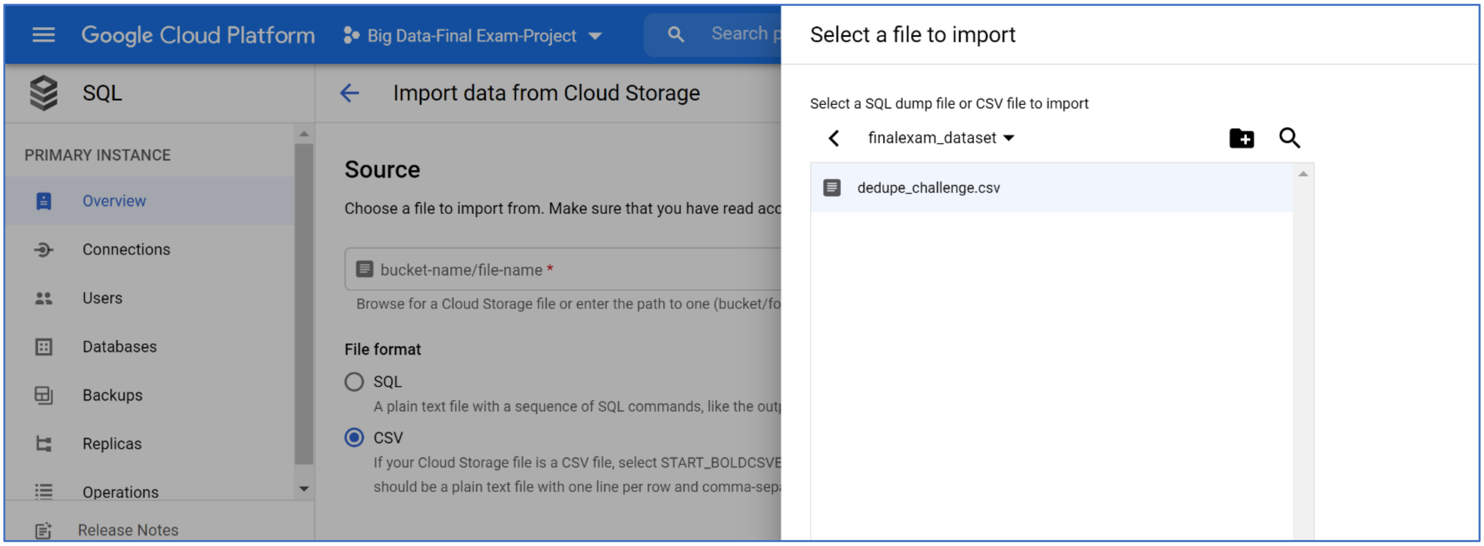


```
mysql> USE wines_db;
Database changed
mysql> CREATE TABLE IF NOT EXISTS Wines (
    -> country varchar(255) ,
    ->  description varchar(255) ,
    ->  designation varchar(255) ,
    ->  points varchar(255) ,
    ->  price varchar(255) ,
    ->  province varchar(255) ,
    ->  region_1 varchar(255) ,
    ->  region_2 varchar(255),
    ->  taster_name varchar(255),
    ->  taster_twitter_handle varchar(255),
    ->  title varchar(255),
    ->  variety varchar(255),
    ->  winery varchar(255));
Query OK, 0 rows affected (0.10 sec)
```

Fig. 5

Load data from Cloud Storage to Cloud SQL

The bucket ‘finalexam_dataset’ that was created earlier contains ‘dedupe_challenge.csv’. (refer Fig. 6)



Google Cloud Platform
Big Data-Final Exam-Project
Search p

SQL

PRIMARY INSTANCE
Overview
Connections
Users
Databases
Backups
Replicas
Operations
Release Notes

Import data from Cloud Storage

Source

Choose a file to import from. Make sure that you have read acc

bucket-name/file-name *

Browse for a Cloud Storage file or enter the path to one (bucket/fo

File format

○ SQL
A plain text file with a sequence of SQL commands, like the out

● CSV
If your Cloud Storage file is a CSV file, select START_BOLDCSVE
should be a plain text file with one line per row and comma-sep

Select a file to import

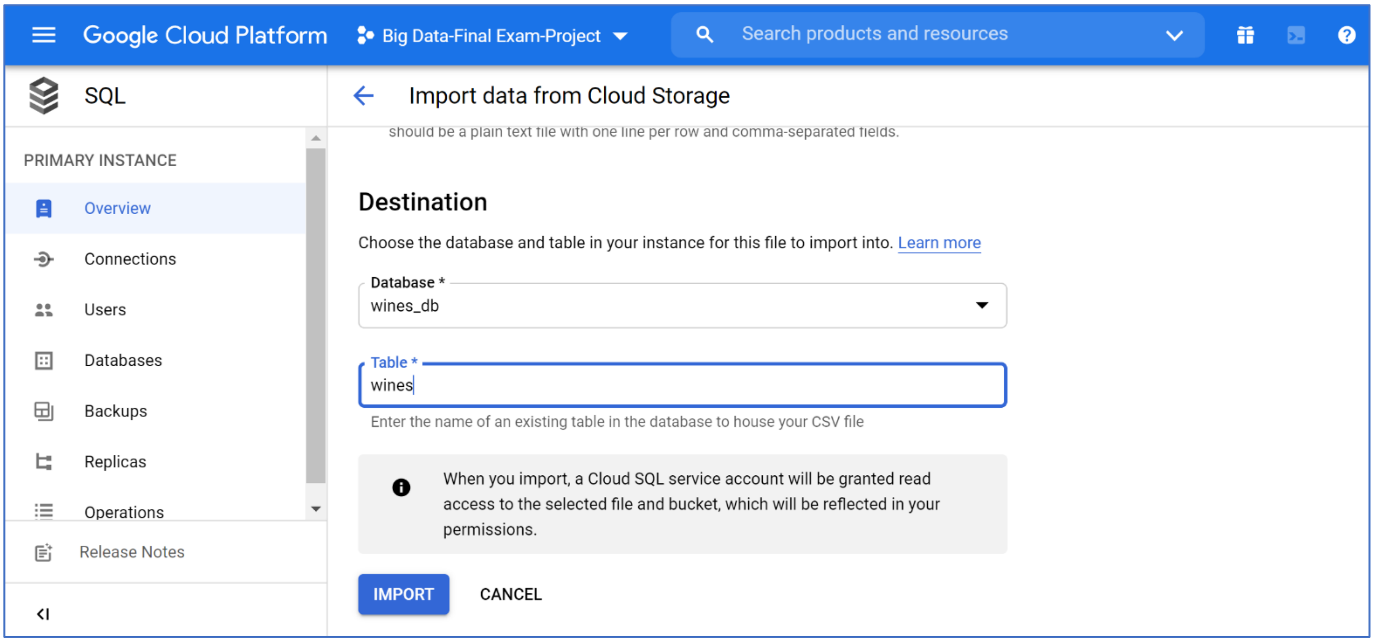
Select a SQL dump file or CSV file to import

finalexam_dataset

dedupe_challenge.csv

Fig. 6

Choose database as ‘wines_db’, table as ‘Wines’. Import to Cloud SQL. (refer Fig. 7)



Google Cloud Platform
Big Data-Final Exam-Project
Search products and resources

SQL

PRIMARY INSTANCE
Overview
Connections
Users
Databases
Backups
Replicas
Operations
Release Notes

Import data from Cloud Storage

should be a plain text file with one line per row and comma-separated fields.

Destination

Choose the database and table in your instance for this file to import into. Learn more

Database *
wines_db

Table *
wines

Enter the name of an existing table in the database to house your CSV file

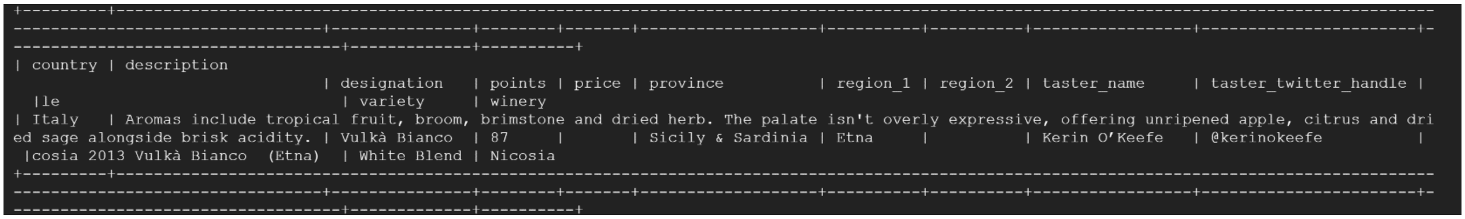
When you import, a Cloud SQL service account will be granted read access to the selected file and bucket, which will be reflected in your permissions.

IMPORT CANCEL

Fig. 7

Checking if dataset from Cloud Storage has been loaded to ‘Wines’ table in ‘wines_db’ database. Query to load one record. It has been loaded one record successfully. (refer Fig. 8)

<u>SQL CODE:</u> `SELECT * FROM Wines LIMIT 1;`



```
+---------+-------------------------------------------------------------------------------------------------------------------------------------------
-------------------------------+--------------+--------+-------+------------------+----------+----------+-----------------+----------------------+-
---------------------------------+------------+----------+
| country | description
                                 | designation   | points | price | province          | region_1 | region_2 | taster_name     | taster_twitter_handle |
  |le                               | variety     | winery
| Italy   | Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dri
ed sage alongside brisk acidity. | Vulkà Bianco  | 87     |       | Sicily & Sardinia | Etna     |          | Kerin O’Keefe   | @kerinokeefe          |
 |cosia 2013 Vulkà Bianco  (Etna)  | White Blend | Nicosia
+---------+-------------------------------------------------------------------------------------------------------------------------------------------
-------------------------------+--------------+--------+-------+------------------+----------+----------+-----------------+----------------------+-
---------------------------------+------------+----------+
```

Fig. 8

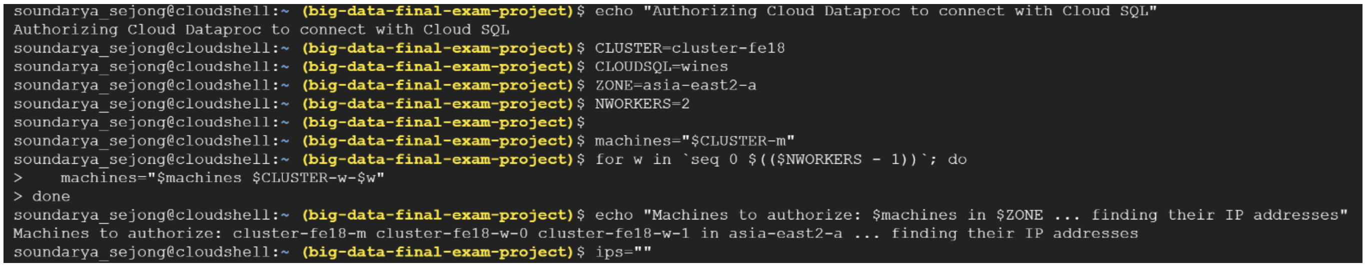
<u>Launch DataProc</u>

Open DataProc which is used for Spark, Hadoop. Create a cluster, it will randomly set a name. “`cluster-fe18`” was created.

Open Cloud Shell to type the following commands to connect to ‘wines’ instance in Cloud SQL (refer Fig. 9, Fig. 10):

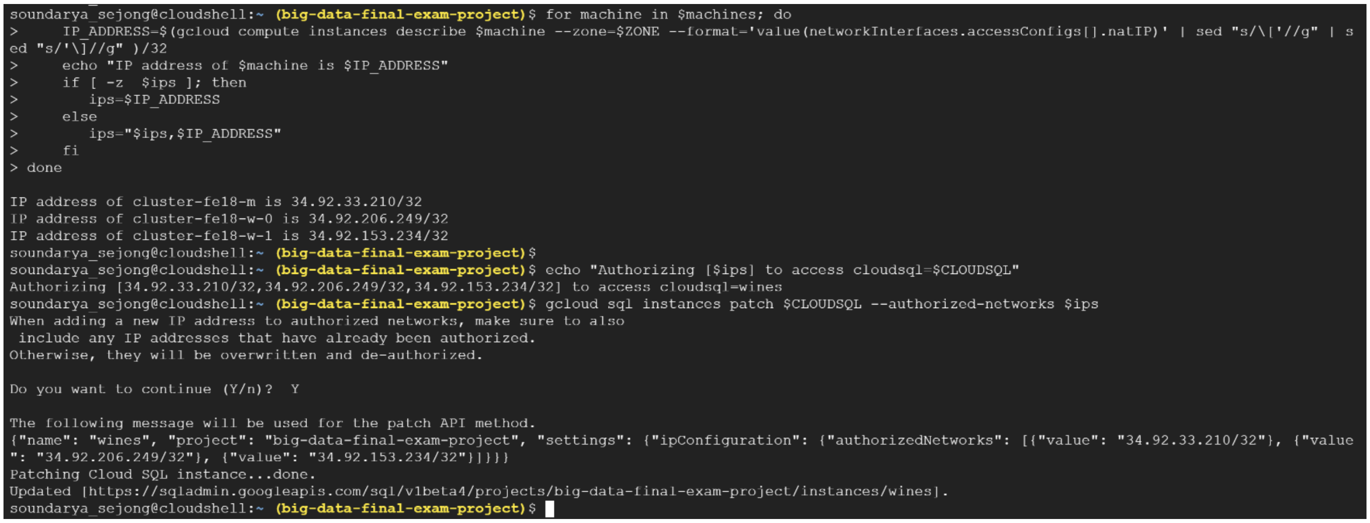
```
echo "Authorizing Cloud Dataproc to connect with Cloud SQL"

CLUSTER=cluster-fe18

CLOUDSQL=wines

ZONE=asia-east2-a

NWORKERS=2


machines="$CLUSTER-m"

for w in `seq 0 $(($NWORKERS - 1))`; do

   machines="$machines $CLUSTER-w-$w"

done


echo "Machines to authorize: $machines in $ZONE ... finding their IP
addresses"

ips=""

for machine in $machines; do
```

```
    IP_ADDRESS=$(gcloud compute instances describe $machine --zone=$ZONE --format='value(networkInterfaces.accessConfigs[].natIP)' | sed "s/\['//g" | sed "s/'\]//g" )/32

    echo "IP address of $machine is $IP_ADDRESS"

    if [ -z  $ips ]; then

       ips=$IP_ADDRESS

    else

       ips="$ips,$IP_ADDRESS"

    fi

done


echo "Authorizing [$ips] to access cloudsql=$CLOUDSQL"

gcloud sql instances patch $CLOUDSQL --authorized-networks $ips
```



```
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ echo "Authorizing Cloud Dataproc to connect with Cloud SQL"
Authorizing Cloud Dataproc to connect with Cloud SQL
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ CLUSTER=cluster-fe18
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ CLOUDSQL=wines
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ ZONE=asia-east2-a
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ NWORKERS=2
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ machines="$CLUSTER-m"
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ for w in `seq 0 $(($NWORKERS - 1))`; do
>    machines="$machines $CLUSTER-w-$w"
> done
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ echo "Machines to authorize: $machines in $ZONE ... finding their IP addresses"
Machines to authorize: cluster-fe18-m cluster-fe18-w-0 cluster-fe18-w-1 in asia-east2-a ... finding their IP addresses
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ ips=""
```

Fig. 9

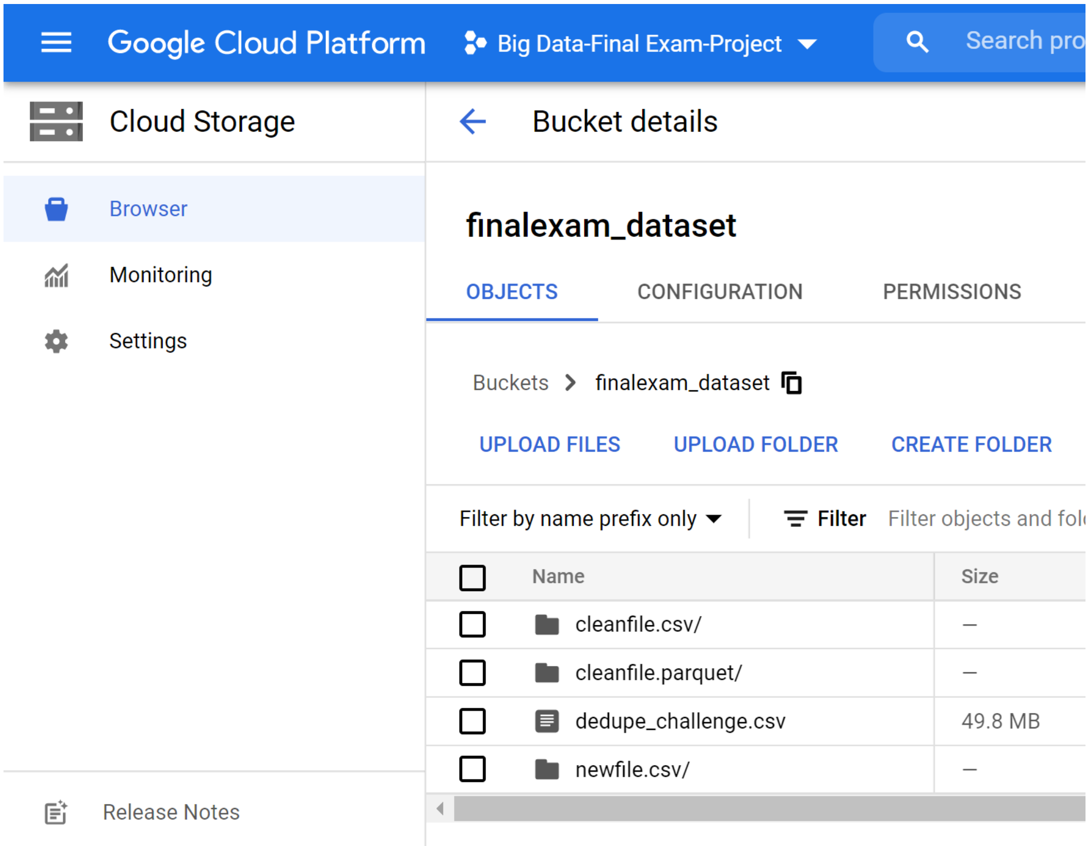


```
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ for machine in $machines; do
>     IP_ADDRESS=$(gcloud compute instances describe $machine --zone=$ZONE --format='value(networkInterfaces.accessConfigs[].natIP)' | sed "s/\['//g" | s
ed "s/'\]//g" )/32
>     echo "IP address of $machine is $IP_ADDRESS"
>     if [ -z  $ips ]; then
>        ips=$IP_ADDRESS
>     else
>        ips="$ips,$IP_ADDRESS"
>     fi
> done

IP address of cluster-fe18-m is 34.92.33.210/32
IP address of cluster-fe18-w-0 is 34.92.206.249/32
IP address of cluster-fe18-w-1 is 34.92.153.234/32
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ echo "Authorizing [$ips] to access cloudsql=$CLOUDSQL"
Authorizing [34.92.33.210/32,34.92.206.249/32,34.92.153.234/32] to access cloudsql=wines
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$ gcloud sql instances patch $CLOUDSQL --authorized-networks $ips
When adding a new IP address to authorized networks, make sure to also
 include any IP addresses that have already been authorized.
Otherwise, they will be overwritten and de-authorized.

Do you want to continue (Y/n)?  Y

The following message will be used for the patch API method.
{"name": "wines", "project": "big-data-final-exam-project", "settings": {"ipConfiguration": {"authorizedNetworks": [{"value": "34.92.33.210/32"}, {"value
": "34.92.206.249/32"}, {"value": "34.92.153.234/32"}]}}}
Patching Cloud SQL instance...done.
Updated [https://sqladmin.googleapis.com/sql/v1beta4/projects/big-data-final-exam-project/instances/wines].
soundarya_sejong@cloudshell:~ (big-data-final-exam-project)$
```

Fig. 10

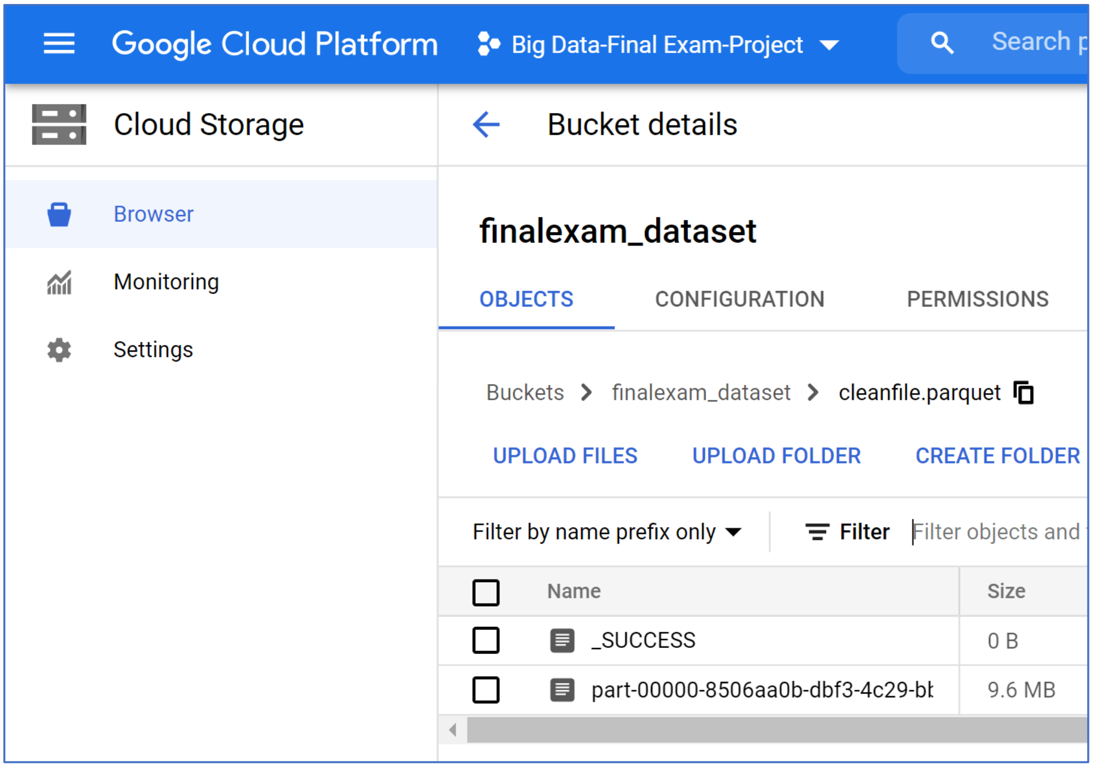
Perform Data Cleaning

Data Cleaning is done on “**Jupyter Notebook**” by launching it from the link given in ‘Component Gateway’ present under ‘Web Instances’ tab in DataProc. The following pages shows the **PySpark 3.1.2** code written Jupyter Notebook used for Final Exam Project. The output screenshots after Data Cleaning is shown. (refer Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15)



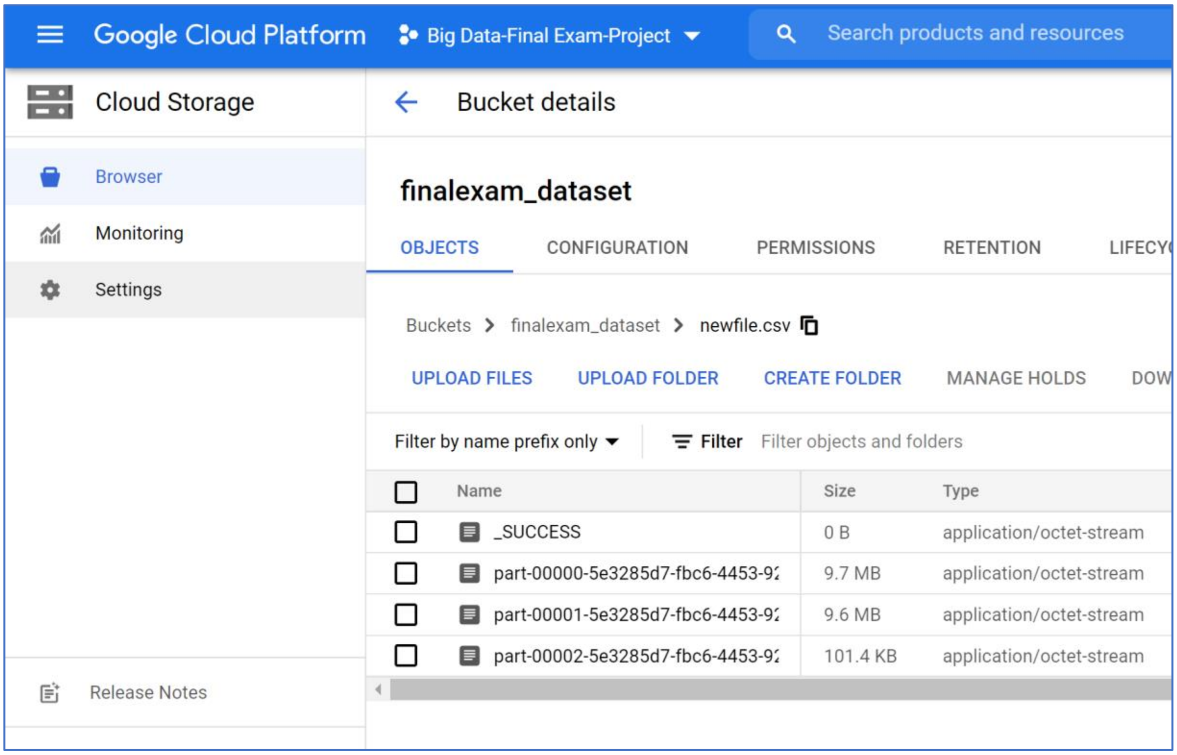
□	Name	Size
□	cleanfile.csv/	–
□	cleanfile.parquet/	–
□	dedupe_challenge.csv	49.8 MB
□	newfile.csv/	–

Fig. 11: Contents in ‘finalexam_dataset’ bucket



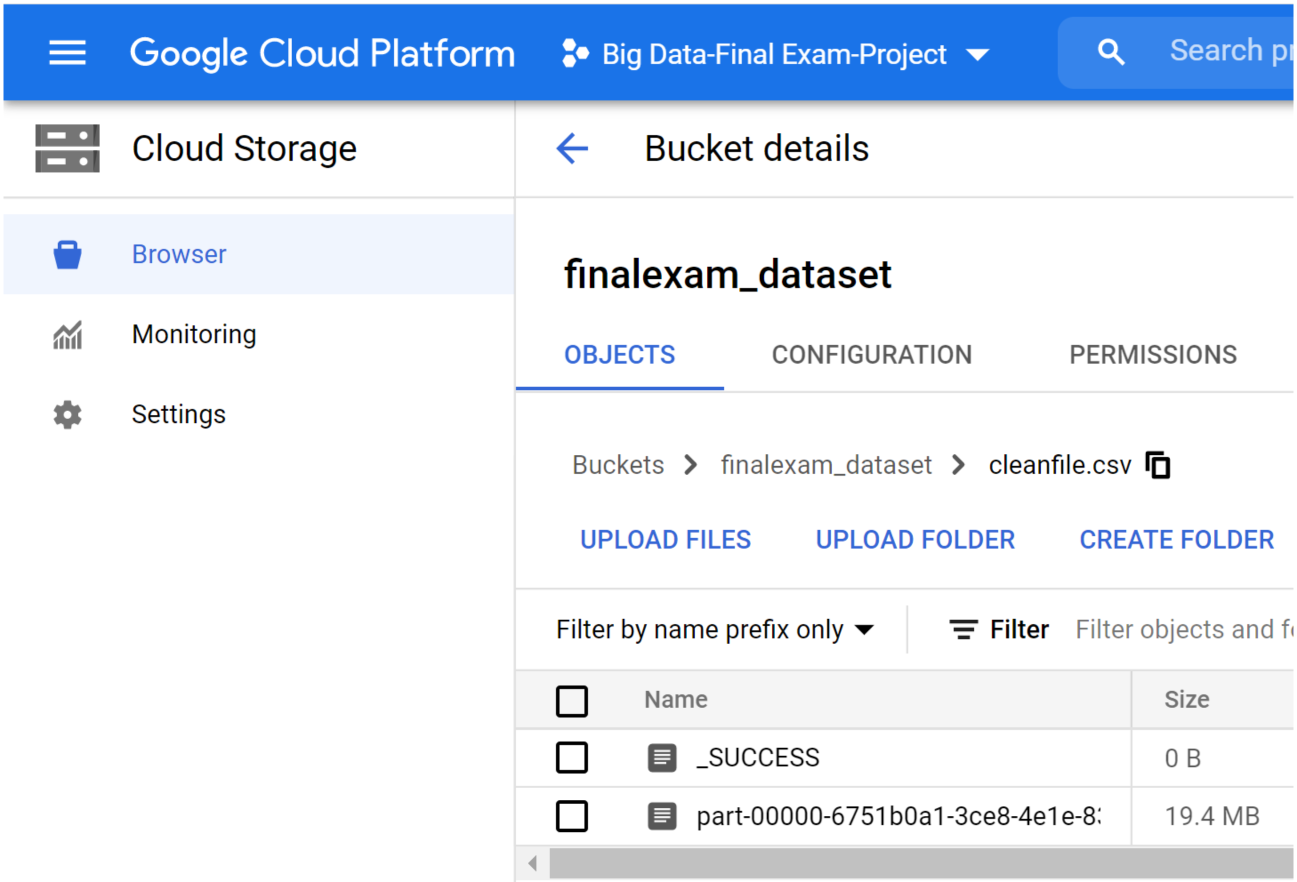
Name	Size
_SUCCESS	0 B
part-00000-8506aa0b-dbf3-4c29-b	9.6 MB

Fig. 12: Contents in cleanfile.parquet



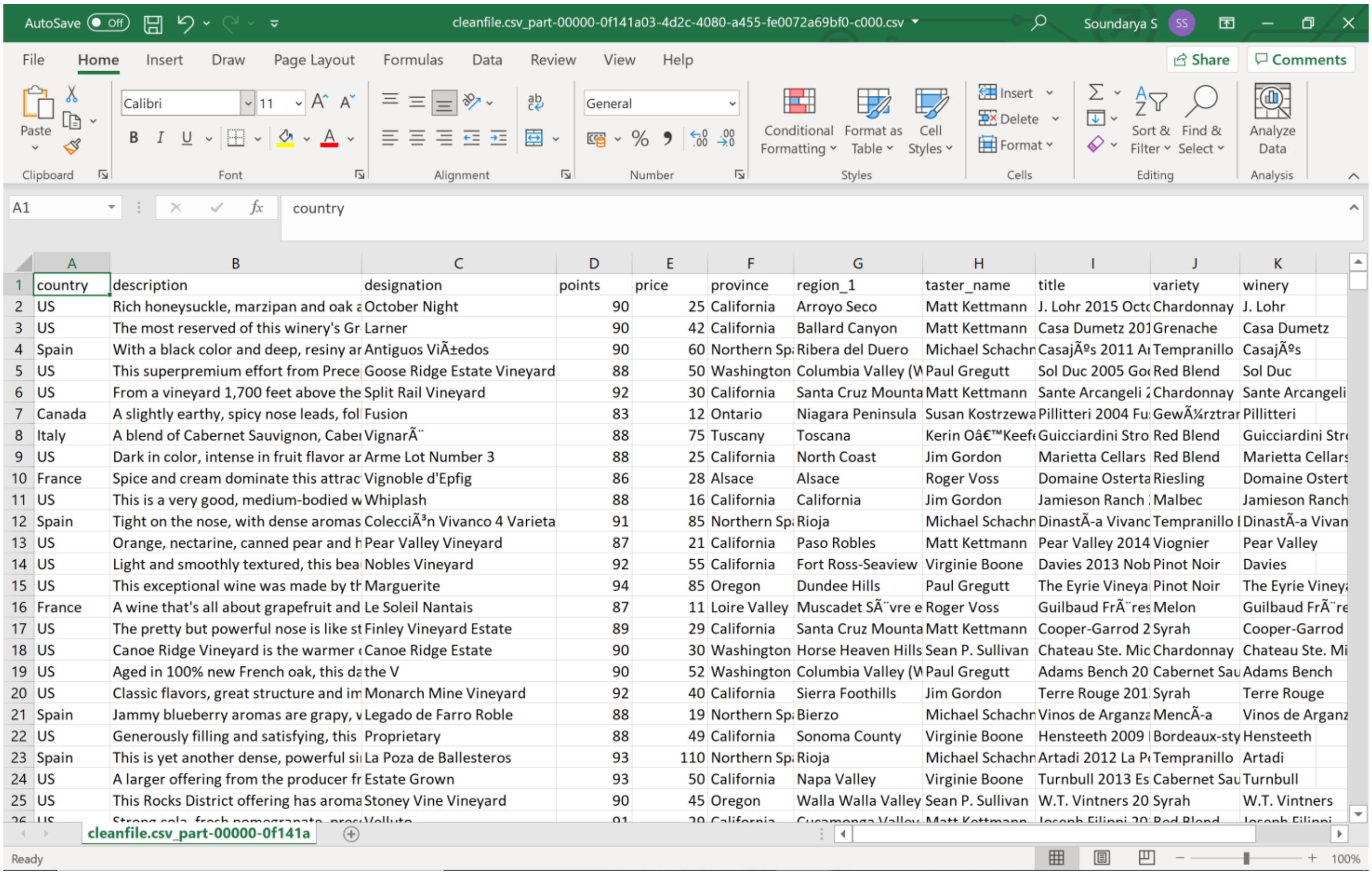
Name	Size	Type
_SUCCESS	0 B	application/octet-stream
part-00000-5e3285d7-fbc6-4453-9	9.7 MB	application/octet-stream
part-00001-5e3285d7-fbc6-4453-9	9.6 MB	application/octet-stream
part-00002-5e3285d7-fbc6-4453-9	101.4 KB	application/octet-stream

Fig. 13: Contents in newfile.csv (to illustrate how GCP partitions files to improve performance and bypass RAM issues)



□	Name	Size
□	_SUCCESS	0 B
□	part-00000-6751b0a1-3ce8-4e1e-8:	19.4 MB

Fig. 14: Contents in cleanfile.csv (to illustrate coalescing given in last code snippet in Jupyter Notebook to save data in one file)



	A	B	C	D	E	F	G	H	I	J	K
1	country	description	designation	points	price	province	region_1	taster_name	title	variety	winery
2	US	Rich honeysuckle, marzipan and oak a	October Night	90	25	California	Arroyo Seco	Matt Kettmann	J. Lohr 2015 Octo	Chardonnay	J. Lohr
3	US	The most reserved of this winery's Gr	Larner	90	42	California	Ballard Canyon	Matt Kettmann	Casa Dumetz 201	Grenache	Casa Dumetz
4	Spain	With a black color and deep, resiny ar	Antiguos ViÃ±edos	90	60	Northern Spa	Ribera del Duero	Michael Schachn	CasajÃºs 2011 Ar	Tempranillo	CasajÃºs
5	US	This superpremium effort from Prece	Goose Ridge Estate Vineyard	88	50	Washington	Columbia Valley (W	Paul Gregutt	Sol Duc 2005 Goo	Red Blend	Sol Duc
6	US	From a vineyard 1,700 feet above the	Split Rail Vineyard	92	30	California	Santa Cruz Mounta	Matt Kettmann	Sante Arcangeli 2	Chardonnay	Sante Arcangeli
7	Canada	A slightly earthy, spicy nose leads, fol	Fusion	83	12	Ontario	Niagara Peninsula	Susan Kostrzewa	Pillitteri 2004 Fu	GewÃ¼rztrar	Pillitteri
8	Italy	A blend of Cabernet Sauvignon, Cabe	VignarÃ¨	88	75	Tuscany	Toscana	Kerin Oâ€™Keefe	Guicciardini Stro	Red Blend	Guicciardini Str
9	US	Dark in color, intense in fruit flavor ar	Arme Lot Number 3	88	25	California	North Coast	Jim Gordon	Marietta Cellars	Red Blend	Marietta Cellars
10	France	Spice and cream dominate this attrac	Vignoble d'Epfig	86	28	Alsace	Alsace	Roger Voss	Domaine Osterta	Riesling	Domaine Ostert
11	US	This is a very good, medium-bodied w	Whiplash	88	16	California	California	Jim Gordon	Jamieson Ranch 2	Malbec	Jamieson Ranch
12	Spain	Tight on the nose, with dense aromas	ColecciÃ³n Vivanco 4 Varieta	91	85	Northern Spa	Rioja	Michael Schachn	DinastÃ-a Vivanc	Tempranillo I	DinastÃ-a Vivan
13	US	Orange, nectarine, canned pear and h	Pear Valley Vineyard	87	21	California	Paso Robles	Matt Kettmann	Pear Valley 2014	Viognier	Pear Valley
14	US	Light and smoothly textured, this bea	Nobles Vineyard	92	55	California	Fort Ross-Seaview	Virginie Boone	Davies 2013 Nob	Pinot Noir	Davies
15	US	This exceptional wine was made by th	Marguerite	94	85	Oregon	Dundee Hills	Paul Gregutt	The Eyrie Vineya	Pinot Noir	The Eyrie Viney
16	France	A wine that's all about grapefruit and	Le Soleil Nantais	87	11	Loire Valley	Muscadet SÃ¨vre e	Roger Voss	Guilbaud FrÃ¨res	Melon	Guilbaud FrÃ¨re
17	US	The pretty but powerful nose is like st	Finley Vineyard Estate	89	29	California	Santa Cruz Mounta	Matt Kettmann	Cooper-Garrod 2	Syrah	Cooper-Garrod
18	US	Canoe Ridge Vineyard is the warmer o	Canoe Ridge Estate	90	30	Washington	Horse Heaven Hills	Sean P. Sullivan	Chateau Ste. Mic	Chardonnay	Chateau Ste. Mi
19	US	Aged in 100% new French oak, this da	the V	90	52	Washington	Columbia Valley (W	Paul Gregutt	Adams Bench 20	Cabernet Sau	Adams Bench
20	US	Classic flavors, great structure and im	Monarch Mine Vineyard	92	40	California	Sierra Foothills	Jim Gordon	Terre Rouge 201	Syrah	Terre Rouge
21	Spain	Jammy blueberry aromas are grapy, w	Legado de Farro Roble	88	19	Northern Spa	Bierzo	Michael Schachn	Vinos de Arganza	MencÃ-a	Vinos de Arganz
22	US	Generously filling and satisfying, this	Proprietary	88	49	California	Sonoma County	Virginie Boone	Hensteeth 2009 I	Bordeaux-sty	Hensteeth
23	Spain	This is yet another dense, powerful si	La Poza de Ballesteros	93	110	Northern Spa	Rioja	Michael Schachn	Artadi 2012 La Po	Tempranillo	Artadi
24	US	A larger offering from the producer fr	Estate Grown	93	50	California	Napa Valley	Virginie Boone	Turnbull 2013 Es	Cabernet Sau	Turnbull
25	US	This Rocks District offering has aroma	Stoney Vine Vineyard	90	45	Oregon	Walla Walla Valley	Sean P. Sullivan	W.T. Vintners 20	Syrah	W.T. Vintners
26	US	Strong cola, fresh pomegranate, pres	Vellute	91	29	California	Cucamonga Valley	Matt Kettmann	Joseph Filippi 20	Red Blend	Joseph Filippi

Fig. 15: After Data Cleaning

Advanced Big Data: Final Exam - Project

Name: SARAVANAN SOUNDARYA SUNDARI

ID: 2021120203

Data Cleaning is usually done to prepare clean data for use in data processing pipelines. To give some examples of Data Cleaning would be removing duplicates, reformatting text, performing calculations, removing incomplete data, removing duplicates, etc.

This Project aims to implement data cleaning on the given "dedupe_challenge.csv" dataset, which was already provided in Midterm Exam to remove duplicates.

Load DataFrame

Reading CSV file

"Cloud Storage" in Google Cloud Platform (GCP) is used to stage data: 'dedupe_challenge.csv'. It is stored under "finalexam_dataset" folder in Cloud Storage.

[1]:

```
df = spark.read.csv("gs://finalexam_dataset/dedupe_challenge.csv", header =
 ↪True, inferSchema = True)
```

Created a database called 'wines_db' in "Cloud SQL" in GCP to store the data from 'dedupe_challenge.csv' by loading data from "Cloud Storage".

Cloud SQL Instances:

[2]:

```
CLOUDSQL_INSTANCE_IP = '34.64.205.24'            #database server IP)
CLOUDSQL_DB_NAME = 'wines_db' #database name
CLOUDSQL_USER = 'root'
CLOUDSQL_PWD  = 'wines'        #database password
```

[3]:

```
#To print Schema
df.printSchema()
```

```
root
 |-- country: string (nullable = true)
 |-- description: string (nullable =true)
 |-- designation: string (nullable =true)
 |-- points: string (nullable = true)
|-- price: string (nullable = true)
|-- province: string (nullable =true)
|-- region_1: string (nullable =true)
|-- region_2: string (nullable =true)
|-- taster_name: string (nullable = true)
|-- taster_twitter_handle: string (nullable = true)
|-- title: string (nullable = true)
|-- variety: string (nullable = true)
|-- winery: string (nullable = true)
```

[4]:
```
df.dtypes
```

[4]:
```
[('country', 'string'), ('description','string'),
   ('designation', 'string'),
   ('points', 'string'),
   ('price', 'string'),
   ('province', 'string'),
   ('region_1', 'string'),
   ('region_2', 'string'),
   ('taster_name', 'string'),
   ('taster_twitter_handle', 'string'), ('title',
   'string'),
   ('variety', 'string'),
   ('winery', 'string')]
```

Counting total no. of records

[5]:
```
df.count()
```

[5]:
```
129984
```

Show top 20 records to check duplicates

[6]:
```
df.show(20)
```

```
+---------+--------------------+--------------------+------+-----+--------------
---+-------------------+-----------------+------------------+-------------------
--+--------------------+------------------+-------------------+
|    country|                  description|                  designation|points|price| province|
     region_1|                  region_2| taster_name|taster_twitter_handle|
     title|                        variety| winery|
+---------+--------------------+--------------------+------+-----+--------------
---+-------------------+-----------------+------------------+-------------------
--+--------------------+------------------+-------------------+
|      Italy|Aromas include tr...|                Vulkà Bianco|        87| null|Sicily &
Sardinia|                   Etna|                    null|        Kerin O’Keefe|
@kerinokeefe|Nicosia 2013 Vulk...|                White Blend|                  Nicosia|
| Portugal|This is ripe and ...|                        Avidagos|        87| 15.0| Douro|
                      null|                null|            Roger Voss|
@vossroger|Quinta dos Avidag...|           Portuguese Red|Quinta dosAvidagos|
|         US|Tart and snappy, ...|                       null|   87|   14.0| Oregon|
          Willamette Valley|Willamette Valley|            Paul Gregutt| @paulgwine
|Rainstorm 2013 Pi...|                      Pinot Gris|                 Rainstorm|
|         US|Pineapple rind, l...|Reserve Late Harvest|                 87| 13.0|
Michigan|Lake Michigan Shore|                        null|Alexander Peartree|
null|St. Julian 2013 R...|              Riesling|   St. Julian|
|         US|Much like the reg...|Vintner's Reserve...|                87| 65.0|
Oregon|     Willamette Valley|Willamette Valley|                  Paul Gregutt|
@paulgwine  |Sweet Cheeks 2012...|                Pinot Noir|              Sweet Cheeks|
|      Spain|Blackberry and ra...|                Ars In Vitro|        87| 15.0| Northern
Spain|                 Navarra|                null| Michael Schachner|
@wineschach|Tandem 2011 Ars I...|Tempranillo-Merlot|                          Tandem|
|      Italy|Here's a bright, ...|                       Belsito|        87| 16.0|Sicily &
Sardinia|                 Vittoria|                  null|        Kerin O’Keefe|
@kerinokeefe|Terre di Giurfo 2...|                      Frappato|       Terre di Giurfo|
|    France|This dry and rest...|                         null|        87| 24.0| Alsace|
    Alsace|                              null|             Roger Voss|
@vossroger|Trimbach 2012 Gew...|          Gewürztraminer|                  Trimbach|
|   Germany|Savory dried thym...|  Shine|                   87| 12.0| Rheinhessen|
   null|                              null|Anna Lee C. Iijima| null|Heinz Eifel 2013 ...|
   Gewürztraminer|                Heinz Eifel|
|    France|This has great de...|                  Les Natures|        87| 27.0| Alsace|
    Alsace|                                   null|                Roger Voss|
@vossroger|Jean-Baptiste Ada...|               Pinot Gris| Jean-Baptiste Adam|
|         US|Soft, supple plum...|             Mountain Cuvée|        87| 19.0|
California|                Napa Valley|                     Napa|       Virginie Boone|
@vboone|Kirkland Signatur...|Cabernet Sauvignon| KirklandSignature|
|    France|This is a dry win...|                          null|        87| 30.0| Alsace|
    Alsace|                              null|             Roger Voss|
@vossroger|Leon Beyer 2012 G...|           Gewürztraminer|                Leon Beyer|
|         US|Slightly reduced,...|                        null|        87| 34.0|
California|          Alexander Valley|                Sonoma|       Virginie Boone|
@vboone|Louis M. Martini ...|Cabernet Sauvignon|    Louis M. Martini|
|      Italy|This is dominated...|                       Rosso|        87| null|Sicily &
Sardinia|                   Etna|                    null|        Kerin O’Keefe|
@kerinokeefe|Masseria Settepor...| Nerello Mascalese|MasseriaSetteporte|
|         US|Building on 150 y...|                        null|        87| 12.0| California|
          Central Coast|                    Central Coast|           Matt Kettmann|
@mattkettmann|Mirassou 2012 Cha...|                 Chardonnay|                 Mirassou|
|   Germany|Zesty orange peel...|                        Devon|        87| 24.0|
Mosel|                      null|                 null|Anna Lee C. Iijima|
null|Richard Böcking 2...|  Riesling|                  Richard Böcking|
|    Germany|Zesty orange peel...|                              Devon|        87| 24.0|
Mosel|                          null|                          null|Anna Lee C. Iijima|
null|Richard Böcking 2...|   Riesling|                       Richard Böcking|
|Argentina|Baked plum, molas...|                                Felix|           87| 30.0| Other|
                                  Cafayate|                       null| Michael Schachner|
@wineschach|Felix Lavaque 201...|                              Malbec|            Felix Lavaque|
|Argentina|Raw black-cherry ...| Winemaker Selection|                        87| 13.0| Mendoza
Province|                       Mendoza|                          null| Michael Schachner|
@wineschach|Gaucho Andino 201...|                              Malbec|            Gaucho Andino|
|        Spain|Desiccated blackb...|Vendimia Seleccio...|                      87| 28.0| Northern
Spain|        Ribera del Duero|                         null| Michael Schachner|
@wineschach|Pradorey 2010 Ven...| Tempranillo Blend|                                   Pradorey|
+---------+--------------------+--------------------+------+-----+--------------
---+-------------------+-----------------+------------------+-------------------
--+--------------------+------------------+-------------------+
only showing top 20 rows
```

Record 15 which contains information about Germany and description “Zesty orange peels...” has a duplicate in Record 16.

Data Cleaning

Dropping duplicates

[7]:

```
#Showing Distinct Records i.e., dropping duplicates from all columns
from pyspark.sql import SparkSession
from pyspark.sql.functions import expr

df1 = df.distinct()
print("No. of Distinct Records:"+str(df1.count())) df1.show(20)
```

```
No. of Distinct Records: 119992
+---------+--------------------+--------------------+------+-----+--------------
---+--------------------+----------------+-----------------+--------------------
-+--------------------+--------------------+--------------------+
|    country|                 description|                    designation|points|price|
province|                       region_1|                 region_2|
taster_name|taster_twitter_handle|                              title|                        variety|
winery|
+---------+--------------------+--------------------+------+-----+--------------
---+--------------------+----------------+-----------------+--------------------
-+--------------------+--------------------+--------------------+
|        Chile|A bright nose wit...|Single Vineyard F...| 87| 18.0|        Leyda Valley|       null|
         null|Michael Schachner| @wineschach|Leyda 2015 Single...|
         Chardonnay|                                       Leyda|
|              US|Rich honeysuckle,...|                October Night|           90| 25.0|
California|                    Arroyo Seco|         Central Coast|            Matt Kettmann|
@mattkettmann|J. Lohr 2015 Octo...|                          Chardonnay|                         J. Lohr|
|            US|Tasty, with pie-f...|                          Reserve|          85| 25.0|
California|                Sonoma Mountain|                   Sonoma|                          null|
null|Work 2004 Reserve...|                              Merlot|                            Work|
|            US|An easy Pinot Noi...|                              null|          87| 28.0|
California|                    Edna Valley|         Central Coast|                          null|
null|Claiborne & Churc...|                      Pinot Noir|Claiborne & Churc...|
|            US|A beautiful spark...|    Ocean Reserve|  92| 40.0| California|       Green
Valley|                     Sonoma|                   null| null|Iron Horse 2007 O...|
             Sparkling Blend|              Iron Horse|
|            US|The most reserved...|                            Larner|          90| 42.0| California|
             Ballard Canyon|                      Central Coast|            Matt Kettmann|
@mattkettmann|Casa Dumetz 2014 ...|                           Grenache|                  Casa Dumetz|
|        Italy|This ruby-hued bl...|Pietralava|             88| null|Sicily & Sardinia|     Etna|
        null|                         Kerin O'Keefe| @kerinokeefe|Antichi Vinai 187...|
                                      Red Blend|             Antichi Vinai 1877|
|        Spain|With a black colo...|                Antiguos Viñedos|              90| 60.0|         Northern
Spain|          Ribera del Duero|                           null|Michael Schachner|
@wineschach|Casajús 2011 Anti...|                      Tempranillo|                          Casajús|
|            US|This superpremium...|Goose Ridge Estat...|                88| 50.0|
Washington|Columbia Valley (WA)| Columbia Valley|                   Paul Gregutt|
@paulgwine  |Sol Duc 2005 Goos...|                          Red Blend|                          Sol Duc|
|            US|A bit of charred ...|                              null|          85| 12.0|
California|                      California|California Other|                  Jim Gordon|
@gordone_cellars|Gnarly Head 2015 ...|                          Pinot Noir|                    Gnarly Head|
|      France|Like many of the ...|                              null|          90|126.0|            Rhône
Valley|                    Hermitage|                    null|        Joe Czerwinski|
@JoeCz|Tardieu-Laurent 2...|                               Syrah|          Tardieu-Laurent|
|            US|Layers of round, ...|                            null|          85| 18.0|
             New York|                     New York|          New York Other|                    Susan
Kostrzewa|  @suskostrzewa|Ventosa 2005 Char...|       Chardonnay|                         Ventosa|
|            US|From a vineyard 1...| Split Rail Vineyard|                  92| 30.0| California|Santa
Cruz Mountains|                             Central Coast|            Matt Kettmann|
@mattkettmann|Sante Arcangeli 2...|                          Chardonnay|          Sante Arcangeli|
|Australia|A soft, medium-bo...|                                    null|          88| 14.0|       Australia
Other|South Eastern Aus...|                           null|        Joe Czerwinski|
@JoeCz|Nugan Family Esta...|          Cabernet Sauvignon|Nugan Family Estates|
|      France|Under the serious...|                              null|     87| 20.0|
Bordeaux|Cadillac Côtes de...|                         null|                Roger Voss|
@vossroger|Château de Lestia...|Bordeaux-style Re...|       Château de Lestiac|
|        Italy|Made from Nerello...|                    Barbazzale|          88| 14.0|Sicily &
Sardinia|                          Etna|                     null|                        null|
null|Cottanera 2011 Ba...|                        Red Blend|                      Cottanera|
| Portugal|Still young, it i...|                                  null| 87| null| Tejo|
                               null|                   null|               Roger Voss|
@vossroger|Quinta do Casal B...|                   Portuguese Red|Quinta do Casal B...|
|      France|From the northern...|                              null|          90| 19.0|
Bordeaux|                    Haut-Médoc|                         null|            Roger Voss|
@vossroger|Château Larrivaux...|Bordeaux-style Re...|              Château Larrivaux|
|            US|This Cab wants a ...|Bell Mountain Vin...|                 90| 52.0|
California|             Alexander Valley|                         Sonoma|                      null|
null|Medlock Ames 2008...|         Cabernet Sauvignon|                 Medlock Ames|
|       Canada|A slightly earthy...|                                     Fusion|          83| 12.0|
Ontario|        Niagara Peninsula|                              null|       Susan Kostrzewa|
@suskostrzewa|Pillitteri 2004 F...|Gewürztraminer-Ri...|                               Pillitteri|
+---------+--------------------+--------------------+------+-----+--------------
---+--------------------+----------------+-----------------+--------------------
-+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Drop Columns

Remove columns that are not required such as ‘region_2’, taster_twitter_handle’.

- region_2 is removed as it is not required, many rows are missing under region_2 resulting in insufficient data.
- ‘taster_twitter_handle’ column which has twitter ids of Wine Tasters is not necessary as there are wine tasters that don’t have twitter handles

[8]:

```
clean1_df =df1.drop('region_2','taster_twitter_handle')
clean1_df.show(5)
```

```
+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+------------
--------+
|country|                 description|                     designation|points|price|               province|
region_1|            taster_name|                              title|                  variety|
winery|
+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+------------
--------+
|     Chile|A bright nose wit...|Single Vineyard F...|              87| 18.0|Leyda Valley|
null|Michael Schachner|Leyda 2015 Single...|                   Chardonnay|
Leyda|
|          US|Rich honeysuckle,...|                   October Night|           90| 25.0|       California|
Arroyo Seco|            Matt Kettmann|J. Lohr 2015 Octo...|                 Chardonnay|
J. Lohr|
|          US|Tasty, with pie-f...|                                Reserve|           85| 25.0|
California|Sonoma Mountain|                                  null|Work 2004 Reserve...|
Merlot|                                  Work|
|          US|An easy Pinot Noi...| null|                              87| 28.0|    California| Edna Valley|
          null|Claiborne & Churc...|                                       Pinot Noir|Claiborne &
Churc...|
|          US|A beautiful spark...|                   Ocean Reserve|           92| 40.0|       California|
Green Valley|                    null|Iron Horse 2007 O...|Sparkling Blend| Iron
Horse|
+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+------------
--------+
only showing top 5 rows
```

Remove Rows

Remove rows for fields that have missing data.

[9]:
```
# Import Pyspark SQL helper functions
from pyspark.sql import functions as F
```

Price

[10]:
```
print("Records where prices are not mentioned: \n")
clean1_df.filter(F.col("price").isNull()).show(5)
```

```
Records where prices are not mentioned:

+--------+--------------------+--------------------+------+-----+---------------
--+--------------------+-------------+--------------------+----------------+----
----------------+
| country|         description|          designation|points|price|       province|
                          region_1| taster_name|                          title|
variety|                      winery|
+--------+--------------------+--------------------+------+-----+---------------
--+--------------------+-------------+--------------------+----------------+----
----------------+
|     Italy|This ruby-hued bl...|  Pietralava|        88| null|Sicily & Sardinia|
     Etna|Kerin O'Keefe|Antichi Vinai 187...|        Red Blend|         Antichi Vinai 1877|
|Portugal|Still young, it i...|                              null|          87| null|
Tejo|                        null|       Roger Voss|Quinta do Casal B...|          Portuguese
Red|Quinta do Casal B...|
|Portugal|This is, as alway...|                  Reserva Branco|          91| null|
Alentejano|                         null|       Roger Voss|Monte daRavasque...|Portuguese
White|Monte da Ravasqueira|
|   France|This wine, solid ...| Domaine de Lavernée|                   89| null|
Beaujolais|    Chiroubles|  Roger Voss|Georges Duboeuf 2...| Gamay|     Georges
Duboeuf|
|     Italy|Almond blossom, s...|Extra Dry Partice...|                  90| null| Veneto|Prosecco di
Valdo...|                                  null|Sorelle Bronca NV...|                 Prosecco|
Sorelle Bronca|
+--------+--------------------+--------------------+------+-----+---------------
--+--------------------+-------------+--------------------+----------------+----
```

```
----------------+
only showing top 5 rows
```

[11]:

```
print("Records where prices are mentioned and storing them in a newdataframe␣
↪'clean2_df': \n")
clean2_df = clean1_df.where(F.col("price").isNotNull())

clean2_df.show(5)
```

```
Records where prices are mentioned and storing them in a newdataframe 'clean2_df':

+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+------------
--------+
|country|         description|                 designation|points|price|    province|
region_1|       taster_name|                   title|          variety|
winery|
+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+------------
--------+
|   Chile|A bright nose wit...|Single Vineyard F...|          87| 18.0|Leyda Valley|
null|Michael Schachner|Leyda 2015 Single...|          Chardonnay|
Leyda|
|      US|Rich honeysuckle,...|          October Night|      90| 25.0|    California|
Arroyo Seco|       Matt Kettmann|J. Lohr 2015 Octo...|      Chardonnay|
J. Lohr|
|      US|Tasty, with pie-f...|                Reserve|      85| 25.0|
California|Sonoma Mountain|                 null|Work 2004 Reserve...|
Merlot|                      Work|
|      US|An easy Pinot Noi...| null|            87| 28.0|  California| Edna Valley|
      null|Claiborne & Churc...|                   Pinot Noir|Claiborne &
Churc...|
|      US|A beautiful spark...|  Ocean Reserve|           92| 40.0|    California|
Green Valley|                 null|Iron Horse 2007 O...|Sparkling Blend|
Iron Horse|
+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+------------
--------+
only showing top 5 rows
```

[12]:

```
clean2_df.count()
```

[12]: 111594

After removing records where prices were not mentioned there are a total of 111,594 records.

Points

```
[13]: print("Records where 'points' is mentioned: \n") clean2_df =
clean2_df.filter(F.col("points").isNotNull())

clean2_df.show()
```

```
Records where 'points' is mentioned:

+---------+--------------------+--------------------+------+-----+--------------
---+--------------------+-----------------+--------------------+----------------
----+--------------------+
|  country|         description|         designation|points|price|
province|            region_1|      taster_name|               title|
variety|              winery|
+---------+--------------------+--------------------+------+-----+--------------
---+--------------------+-----------------+--------------------+----------------
----+--------------------+
|    Chile|A bright nose wit...|Single Vineyard F...|    87| 18.0|      Leyda
Valley|                null|Michael Schachner|Leyda 2015 Single...| Chardonnay|
                   Leyda|
|       US|Rich honeysuckle,...|       October Night|    90| 25.0|
California|         Arroyo Seco|   Matt Kettmann|J. Lohr 2015 Octo...|
Chardonnay|             J. Lohr|
|       US|Tasty, with pie-f...|             Reserve|    85| 25.0|
California|     Sonoma Mountain|             null|Work 2004 Reserve...|
Merlot|                    Work|
|       US|An easy Pinot Noi...|                null|    87| 28.0|
California|         Edna Valley|             null|Claiborne & Churc...|
Pinot Noir|Claiborne & Churc...|
|       US|A beautiful spark...|       Ocean Reserve|    92| 40.0|
California|        Green Valley|             null|Iron Horse 2007 O...|
Sparkling Blend|          Iron Horse|
|       US|The most reserved...|              Larner|    90| 42.0|
California|      Ballard Canyon|   Matt Kettmann|Casa Dumetz 2014 ...|
Grenache|         Casa Dumetz|
|    Spain|With a black colo...|     Antiguos Viñedos|    90| 60.0|     Northern
Spain|     Ribera del Duero|Michael Schachner|Casajús 2011 Anti...| Tempranillo|  Casajús|
|       US|This superpremium...|Goose Ridge Estat...|    88| 50.0|
Washington|Columbia Valley (WA)|     Paul Gregutt|Sol Duc 2005 Goos...| Red
Blend|   Sol Duc|
|       US|A bit of charred ...|                null|    85| 12.0|
California|          California|      Jim Gordon|Gnarly Head 2015 ...|
Pinot Noir|         Gnarly Head|
|   France|Like many of the ...|                null|    90|126.0|        Rhône
Valley|            Hermitage|   Joe Czerwinski|Tardieu-Laurent 2...|
Syrah|       Tardieu-Laurent|
|           US|Layers of round, ...|    null|                     85| 18.0|    New York|
            New York|  Susan Kostrzewa|Ventosa 2005 Char...|
Chardonnay|                   Ventosa|
|           US|From a vineyard 1...| Split Rail Vineyard|                    92| 30.0|
California|Santa Cruz Mountains|             Matt Kettmann|Sante Arcangeli 2...|
Chardonnay|          Sante Arcangeli|
|Australia|A soft, medium-bo...|    null|                     88| 14.0|              Australia
Other|South Eastern Aus...|          Joe Czerwinski|Nugan Family Esta...|            Cabernet
Sauvignon|Nugan Family Estates|
|     France|Under the serious...|                           null|         87| 20.0|
Bordeaux|Cadillac Côtes de...|                  Roger Voss|Château deLestia...|Bordeaux- style
Re...|           Château de Lestiac|
|       Italy|Made from Nerello...|                   Barbazzale|          88| 14.0|Sicily &
Sardinia|                           Etna|                  null|Cottanera 2011 Ba...|
Red Blend|                   Cottanera|
|     France|From the northern...|                           null|         90| 19.0|
Bordeaux|                  Haut-Médoc|           Roger Voss|ChâteauLarrivaux...|Bordeaux- style
Re...|              Château Larrivaux|
|           US|This Cab wants a ...|Bell Mountain Vin...|     90| 52.0| California|
            Alexander Valley|          null|Medlock Ames 2008...| Cabernet Sauvignon|
            Medlock Ames|
|     Canada|A slightly earthy...|                          Fusion|         83| 12.0|
Ontario|        Niagara Peninsula|          Susan Kostrzewa|Pillitteri 2004
F...|Gewürztraminer-Ri...|                     Pillitteri|
|       Italy|A blend of Cabern...|                        Vignarè|          88| 75.0|
Tuscany|                       Toscana|        Kerin O’Keefe|GuicciardiniStro...| Red
Blend|Guicciardini Strozzi|
|           US|Dark in color, in...|          Arme Lot Number 3|          88| 25.0|
California|                  North Coast|                Jim Gordon|Marietta Cellars ...| Red
Blend|      Marietta Cellars|
+---------+--------------------+--------------------+------+-----+--------------
---+--------------------+-----------------+--------------------+----------------
----+--------------------+
only showing top 20 rows
```

[14]:
```
clean2_df.count()
```

[14]: 111590

After removing records where the points for wine are not mentioned there are a total of 111,590 records.

Determine rows that converts properly to an integer value

Removing rows for column such as ‘points’, ‘prices’, that do not convert properly to an integer value as there might be data where it doesn’t contain integer values.

Points

[15]:

```
print("Total no. of records: " + str(clean2_df.count())) print("\nNo. of rows that
converts properly to an integer value:")
clean2_df =clean2_df.filter(F.col('points').cast("int").isNotNull())
clean2_df.filter(F.col('points').cast("int").isNotNull()).count()
```

```
Total no. of records: 111590

No. of rows that converts properly to an integervalue: [15]: 111570
```

Price

[16]:

```
print("Total no. of records: " + str(clean2_df.count())) print("\nNo. of rows that
converts properly to an integer value: ") clean2_df =
clean2_df.filter(F.col('price').cast("int").isNotNull())
clean2_df.filter(F.col('price').cast("int").isNotNull()).count()
```

```
Total no. of records: 111570

No. of rows that converts properly to an integervalue: [16]: 111570
```

Country

[17]:

```
print("Records where 'country' is mentioned: \n") clean2_df =
clean2_df.filter(F.col("country").isNotNull()) clean2_df.show(5)
```

```
Records where 'country' is mentioned:

+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+------------
--------+
|country|          description|               designation|points|price|          province|
region_1|         taster_name|                    title|            variety|
winery|
+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+------------
--------+
|    Chile|A bright nose wit...|Single Vineyard F...|              87| 18.0|Leyda Valley|
null|Michael Schachner|Leyda 2015 Single...|                   Chardonnay|
Leyda|
|          US|Rich honeysuckle,...|                October Night|          90| 25.0|      California|
Arroyo Seco|          Matt Kettmann|J. Lohr 2015 Octo...|              Chardonnay|
J. Lohr|
|       US|Tasty, with pie-f...|                  Reserve|      85| 25.0|
California|Sonoma Mountain|                 null|Work 2004 Reserve...|
Merlot|                   Work|
|       US|An easy Pinot Noi...| null|                  87| 28.0|   California| Edna Valley|
       null|Claiborne & Churc...|                      Pinot Noir|Claiborne &
Churc...|
|       US|A beautiful spark...|  Ocean Reserve|          92| 40.0|     California|
Green Valley|                 null|Iron Horse 2007 O...|Sparkling Blend|
Iron Horse|
+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+------------
--------+
only showing top 5 rows
```

[18]:

```
clean2_df.count()
```

[18]: 111515

After removing records where the place of origin/country is not mentioned there are a total of 111,515 records.

Designation

[19]:

```
print("Records where 'designation' is not mentioned: \n") clean2_df =
clean2_df.filter(F.col("designation").isNotNull()) clean2_df.show(5)
```

```
Records where 'designation' is not mentioned:

+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+-----------+
|country|          description|              designation|points|price|          province|
region_1|         taster_name|                     title|          variety|        winery|
+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+-----------+
|   Chile|A bright nose wit...|Single Vineyard F...|          87| 18.0|Leyda Valley|
null|Michael Schachner|Leyda 2015 Single...|                Chardonnay|       Leyda|
|       US|Rich honeysuckle,...|            October Night|      90| 25.0|     California|
Arroyo Seco|        Matt Kettmann|J. Lohr 2015 Octo...|            Chardonnay|     J. Lohr|
|       US|Tasty, with pie-f...|                  Reserve|      85| 25.0|
California|Sonoma Mountain|                 null|Work 2004 Reserve...|
Merlot|          Work|
|       US|A beautiful spark...|  Ocean Reserve|          92| 40.0|     California| Green
Valley| null|Iron Horse 2007 O...|Sparkling Blend| IronHorse|
|       US|The most reserved...|                   Larner|      90| 42.0|     California|
Ballard Canyon|          Matt Kettmann|Casa Dumetz 2014 ...|           Grenache|Casa
```

```
Dumetz|
+-------+--------------------+--------------------+------+-----+------------+---
------------+-----------------+--------------------+---------------+-----------+
only showing top 5 rows
```

[20]:

```
clean2_df.count()
```

[20]: 79470

After removing records where the designation is not mentioned there are a total of 79,470 records.

Region_1

[21]:

```
print("Records where 'region_1' is mentioned: \n") clean2_df =
clean2_df.filter(F.col("region_1").isNotNull()) clean2_df.show(5)
```

```
Records where 'region_1' is mentioned:

+-------+--------------------+----------------+------+-----+--------------+-----
-----------+-----------------+--------------------+---------------+-----------+
|country|                 description|            designation|points|price| province| region_1|
                    taster_name|                              title|                variety|            winery|
+-------+--------------------+---------------+------+-----+-------------+-----
-----------+-----------------+--------------------+---------------+-----------+
|        US|Rich honeysuckle,...|            October Night|           90| 25.0|         California| Arroyo
Seco|   Matt Kettmann|J. Lohr 2015 Octo...|                                    Chardonnay|          J. Lohr|
|        US|Tasty, with pie-f...|                           Reserve|          85| 25.0|         California|
Sonoma Mountain|                           null|Work 2004 Reserve...|                             Merlot|
Work|
|        US|A beautiful spark...|    Ocean Reserve|                      92| 40.0|         California| Green
Valley| null|Iron Horse 2007 O...|Sparkling Blend| IronHorse|
|        US|The most reserved...|                          Larner|          90| 42.0|         California|
Ballard Canyon|            Matt Kettmann|Casa Dumetz 2014 ...| Grenache|Casa Dumetz|
|   Spain|With a black colo...|Antiguos Viñedos|                      90| 60.0|Northern Spain|Ribera
del Duero|Michael Schachner|Casajús 2011 Anti...|                Tempranillo| Casajús|
+-------+--------------------+---------------+------+-----+-------------+-----
-----------+-----------------+--------------------+---------------+-----------+
only showing top 5 rows
```

[22]:

```
clean2_df.count()
```

[22]: 64754

After removing records where region_1 is not mentioned there are a total of 64,754 records.

Taster’s Name

[23]:
```
print("Records where 'taster_name' is mentioned: \n") clean2_df =
clean2_df.filter(F.col("taster_name").isNotNull()) clean2_df.show(5)
```

```
Records where 'taster_name' is mentioned:

+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
|country|          description|               designation|points|price|          province|
region_1|       taster_name|                    title|         variety|              winery|
+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
|      US|Rich honeysuckle,...|          October Night|      90| 25.0|      California|
Arroyo Seco|       Matt Kettmann|J. Lohr 2015 Octo...| Chardonnay|                J. Lohr|
|      US|The most reserved...|                   Larner|      90| 42.0|      California|
Ballard Canyon|         Matt Kettmann|Casa Dumetz 2014 ...|         Grenache|       Casa
Dumetz|
|   Spain|With a black colo...|        Antiguos Viñedos|      90| 60.0|Northern Spain|
Ribera del Duero|Michael Schachner|Casajús 2011 Anti...|Tempranillo|
Casajús|
|      US|This superpremium...|Goose Ridge Estat...|          88| 50.0|
Washington|Columbia Valley (WA)|               Paul Gregutt|Sol Duc 2005 Goos...|          Red
Blend|  Sol Duc|
|      US|From a vineyard 1...| Split Rail Vineyard|              92| 30.0|
California|Santa Cruz Mountains|               Matt Kettmann|Sante Arcangeli 2...|
Chardonnay|Sante Arcangeli|
+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
only showing top 5 rows
```

[24]:
```
clean2_df.count()
```

[24]: 49541

After removing records where Wine Taster’s name is not mentioned there are a total of 49,541 records.

Title

[25]:
```
print("Records where 'title' is mentioned: \n") clean2_df =
clean2_df.filter(F.col("title").isNotNull()) clean2_df.show(5)
```

```
Records where 'title' is mentioned:

+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
|country|         description|         designation|points|price|      province|
region_1|      taster_name|               title|     variety|        winery|
+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
|     US|Rich honeysuckle,...|       October Night|    90| 25.0|    California|
Arroyo Seco|      Matt Kettmann|J. Lohr 2015 Octo...| Chardonnay|          J. Lohr|
|     US|The most reserved...|              Larner|    90| 42.0|    California|
Ballard Canyon|      Matt Kettmann|Casa Dumetz 2014 ...|     Grenache|     Casa
Dumetz|
|  Spain|With a black colo...|    Antiguos Viñedos|    90| 60.0|Northern Spain|
Ribera del Duero|Michael Schachner|Casajús 2011 Anti...|Tempranillo|
Casajús|
|     US|This superpremium...|Goose Ridge Estat...|      88| 50.0|
Washington|Columbia Valley (WA)|      Paul Gregutt|Sol Duc 2005 Goos...|      Red
Blend|  Sol Duc|
|     US|From a vineyard 1...| Split Rail Vineyard|      92| 30.0|
California|Santa Cruz Mountains|      Matt Kettmann|Sante Arcangeli 2...|
Chardonnay|Sante Arcangeli|
+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
only showing top 5 rows
```

[26]:
```
clean2_df.count()
```

[26]: 49541

In the Dataframe, the field ‘Title’ is not empty/null i.e., all records mention the Title. Therefore, no records are removed. Total = 49,541 records.

Wine Variety

[27]:

```
print("Records where 'variety' is mentioned: \n") clean2_df =
clean2_df.filter(F.col("variety").isNotNull()) clean2_df.show(5)
```

```
Records where 'variety' is mentioned:

+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
|country|                description|                     designation|points|price|                  province|
region_1|             taster_name|                              title|           variety|                    winery|
+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
|          US|Rich honeysuckle,...|                    October Night|           90| 25.0|           California|
Arroyo Seco|            Matt Kettmann|J. Lohr 2015 Octo...| Chardonnay|                        J. Lohr|
|          US|The most reserved...|                              Larner|          90| 42.0|           California|
Ballard Canyon|             Matt Kettmann|Casa Dumetz 2014 ...|             Grenache|         Casa
Dumetz|
|     Spain|With a black colo...|                 Antiguos Viñedos|             90| 60.0|Northern Spain|
Ribera del Duero|Michael Schachner|Casajús 2011 Anti...|Tempranillo|
Casajús|
|          US|This superpremium...|Goose Ridge Estat...|                  88| 50.0|
Washington|Columbia Valley (WA)|                   Paul Gregutt|Sol Duc 2005 Goos...|              Red
Blend|  Sol Duc|
|          US|From a vineyard 1...| Split Rail Vineyard|                    92| 30.0|
California|Santa Cruz Mountains|                     Matt Kettmann|Sante Arcangeli 2...|
Chardonnay|Sante Arcangeli|
+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
only showing top 5 rows
```

[28]:

```
clean2_df.count()
```

[28]: 49541

In the Dataframe, the field ‘Variety’ is not empty/null i.e., all records mention the Wine Variety. Therefore, no records are removed. total = 49,541 records.

Winery

[29]:
```
print("Records where 'winery' is mentioned: \n") clean2_df =
clean2_df.filter(F.col("winery").isNotNull()) clean2_df.show(5)
```

```
Records where 'winery' is mentioned:

+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
|country|          description|          designation|points|price|        province|
region_1|       taster_name|                title|      variety|           winery|
+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
|      US|Rich honeysuckle,...|       October Night|     90| 25.0|      California|
Arroyo Seco|      Matt Kettmann|J. Lohr 2015 Octo...| Chardonnay|          J. Lohr|
|      US|The most reserved...|              Larner|     90| 42.0|      California|
Ballard Canyon|      Matt Kettmann|Casa Dumetz 2014 ...|     Grenache|     Casa
Dumetz|
|   Spain|With a black colo...|     Antiguos Viñedos|     90| 60.0|Northern Spain|
Ribera del Duero|Michael Schachner|Casajús 2011 Anti...|Tempranillo|
Casajús|
|      US|This superpremium...|Goose Ridge Estat...|        88| 50.0|
Washington|Columbia Valley (WA)|        Paul Gregutt|Sol Duc 2005 Goos...|      Red
Blend|  Sol Duc|
|      US|From a vineyard 1...| Split Rail Vineyard|      92| 30.0|
California|Santa Cruz Mountains|         Matt Kettmann|Sante Arcangeli 2...|
Chardonnay|Sante Arcangeli|
+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
only showing top 5 rows
```

[30]:
```
clean2_df.count()
```

[30]: 49541

‘Winery’ field is still mentioned for every record in the dataset, i.e., not empty/null therefore no records are removed therefore, total of 49,541 records.

From a total of 129984 in ‘dedupe_challenge.csv’, after dropping duplicate records, removing columns that are unnecessary, removing rows that have null fields, etc, there is a grand total of 49,541 distinct, unique records prepared for further data cleaning and dataanalysis.

Infering that it had approx. 62% of unclean data.

Therefore, less than half, which is found out to be approx. 38% of data from the original dataset i.e., 49,541 clean, distinct, unique records can be used for further data cleaning and data analysis.

[31]:

```
clean2_df.show(5)
```

```
+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
|country|         description|         designation|points|price|       province|
region_1|      taster_name|               title|    variety|         winery|
+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
|     US|Rich honeysuckle,...|       October Night|    90| 25.0|    California|
Arroyo Seco|     Matt Kettmann|J. Lohr 2015 Octo...| Chardonnay|         J. Lohr|
|     US|The most reserved...|              Larner|    90| 42.0|    California|
Ballard Canyon|     Matt Kettmann|Casa Dumetz 2014 ...|   Grenache|    Casa
Dumetz|
|  Spain|With a black colo...|    Antiguos Viñedos|    90| 60.0|Northern Spain|
Ribera del Duero|Michael Schachner|Casajús 2011 Anti...|Tempranillo|
Casajús|
|     US|This superpremium...|Goose Ridge Estat...|    88| 50.0|
Washington|Columbia Valley (WA)|       Paul Gregutt|Sol Duc 2005 Goos...|      Red
Blend|  Sol Duc|
|     US|From a vineyard 1...| Split Rail Vineyard|    92| 30.0|
California|Santa Cruz Mountains|       Matt Kettmann|Sante Arcangeli 2...|
Chardonnay|Sante Arcangeli|
+-------+--------------------+--------------------+------+-----+--------------+-
-------------------+-----------------+--------------------+-----------+---------
------+
only showing top 5 rows
```

The above clean Dataframe ‘clean2_df’ shows first 5 records.

Save Data

Save data by storing it in file formats for further Data Processing or Data Analysis.

Saving as parquet

For further data processing or data analysis using Spark, save file as ‘Parquet’ because Spark is highly optimized for Parquet files.

[32]:

```
#Saving the parquet file in Cloud Storage
clean2_df.coalesce(1).write.parquet('gs://finalexam_dataset/cleanfile.parquet')
```

The file is stored in “Cloud Storage” where the file is partitioned. Spark generally stores data in separate files to improve performance and bypass RAM issues. Here it is coalesced to one file as the file is not of huge size.

It is allowed to save to other file formats too such .csv file in the next section. It can also be difficult to read Parquet files outside of Spark.

Saving to a new csv file

[33]:

```
clean2_df.write.csv("gs://finalexam_dataset/newfile.csv")
```

The file is stored in “Cloud Storage” where the csv file is partitioned to 3 files for the same reason as mentioned above, i.e., to improve performance and bypass RAM issues.

The dataset is small therefore can bypass those concerns to coalesce save data in one file.

[34]:

```
# Coalesce and save data in CSV format

clean2_df.coalesce(1).write.csv('gs://finalexam_dataset/cleanfile.csv',
 ↪header=True)
```