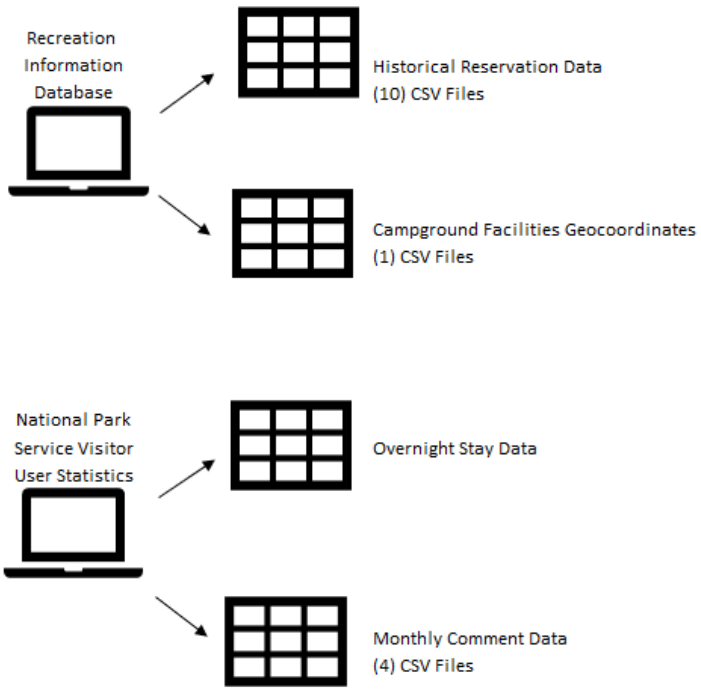
ETL Project Final Report

Jenni Davis, Susan Farago, David Jimenez, Austin Olea, & Elizabeth Conway

With the growing number of apps available to reserve camping spots in Colorado, enthusiasts need a starting point to narrow their search for adventure. Our database identifies Colorado campground locations by latitude and longitude coordinates, provides the number of camping spaces at each location, and lists the allowable equipment to be used at each site. The database also provides enthusiasts with the opportunity to review historical availability from 2010 - 2020 and reservation lead-time by location.

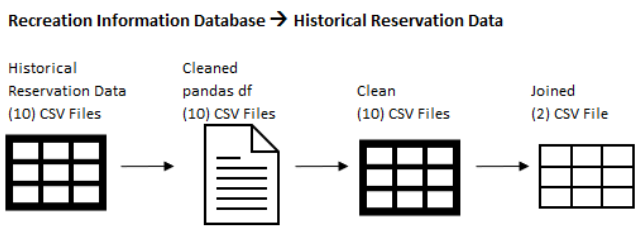
Extract:



Our primary source of data came from the Recreation Information Database (ridb.recreation.gov). Their database provides users with information associated with recreational activities throughout the US. Ten csv files were downloaded for each respective year (2010 – 2020). Each initial file contained 56 columns of attributes associated with a single reservation booked using Recreation.gov. Additional data from the csv files containing the camping facility's latitude and longitude were downloaded for each respective year (2010 – 2020).

Our second source of data came from National Park Service Visitor User Statistics (irma.nps.gov/STATS). Their databases provide a myriad of csv accessible files on recreational activities occurring within National Parks. Four csv files were downloaded for each National Park located in Colorado containing overnight stay data from 2010 – 2020. Monthly visitation comments from 2010 – 2020 were downloaded as csv files, by National Park, respectively.

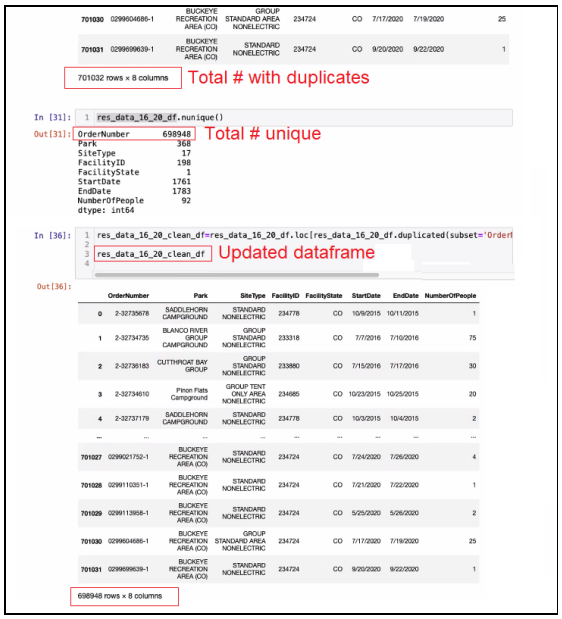
Transform:



The historical reservation data extracted from the Recreation Information Database (RIDB) for 2010 – 2020, respectively, was far too large to be cleaned locally. Our team utilized pandas to load the data into a data frame for cleaning.

The first objective was to rid the files of the columns not relevant to our final database and / or with duplicative information. We first created a new data frame by removing 48 columns not applicable to our database. The remaining table contained the following: OrderNumber, Park, SiteType, FacilityID, FacilityState, StartDate, EndDate, and NumberOfPeople.

Next, we filtered on Colorado within our new data frame using 'FacilityState' / 'CO'. From here we determined that the column 'SiteType' contained unrelated activity bookings. We created a list of desired values and set column 'SiteType' to return only the values present in the list.

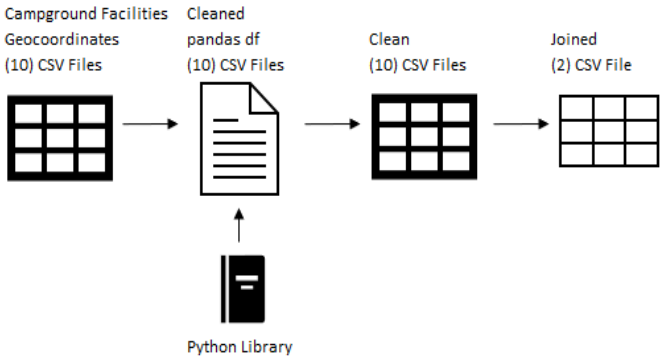


We encountered duplicate OrderNumbers in the historical reservation data. Upon further analysis we decided to drop duplicate data in the OrderNumber column which impacted 2,084 records (see image below).

OrderNumber	Park	SiteType	FacilityID	FacilityState	StartDate	EndDate	NumberOfPeople
2-13888626	GLACIER BASIN CA	GROUP TENT ONLY	232462	CO	8/12/2010	8/13/2010	17
2-13883425	BOGAN FLATS CAM	STANDARD NONELE	232158	CO	9/3/2010	9/6/2010	6
2-13884441	BELLAIRE LAKE CA	STANDARD ELECTF	233137	CO	8/24/2010	8/27/2010	2
2-13887874	GLACIER BASIN CA	STANDARD NONELE	232462	CO	8/7/2010	8/8/2010	1
2-13887885	BUFFALO CAMPGR	STANDARD NONELE	231875	CO	8/13/2010	8/15/2010	6
2-13883441	ASPENGLEN CAMP	STANDARD NONELE	233187	CO	9/17/2010	9/19/2010	4

Once completed, our data frames were checked, confirmed, and exported to csv files, respectively. The ten csv files were joined locally into two csv files.

Recreation Information Database → Campground Facilities Geocoordinates



Additional data of interest present in the RIDB reservation data included the facility coordinates. The Python library pandas and reverse-geocode were used to transform and acquire desired information for an additional table to be included in the final database.

To access the lat/lon coordinates of each campground we selected the following columns from the RIDB reservation csv files and created a new dataframe using pandas library: RegionDescription, Park, FacilityID, FacilityState, FacilityLongitude, and FacilityLatitude.

The RIDB reservation data did not include city or county location associated with the facilities lat/lon coordinates. To allow for queries on campground location proximity within the state we utilized the python library reverse-geocode to append city, state, and county information to the data frame described above. From this data frame we dropped and renamed columns for our final table: RegionDescription, Park, FacilityID, FacilityState, FacilityLongitude, FacilityLatitude, CityPlace, County.

Geocode Data

```
In [21]:
#Read geocode data and load as dataframe
geo_file= os.path.join("Data","Reverse_GeocodeInfo.csv")
geo_df = pd.read_csv(geo_file)
geo_df
```

3	Comanche RD-FS	Picket Wire Canyonlands Guided Auto Tour	234166.0	CO	-103.524	37.9857	La Junta	Otero County
4	Great Sand Dunes National Park	Pinon Flats Campground	234685.0	CO	-105.517	37.7333	Westcliffe	Custer County
...	...	...	...	...	...	...	...	...
195	SAN CARLOS RD -FS	DAVENPORT CAMPGROUND	234684.0	CO	-105.070	38.0556	Colorado City	Pueblo County
196	GRAND VALLEY RD-FS	BLACK BEAR	233216.0	CO	-108.081	39.0522	Cedaredge	Delta County

197 rows × 8 columns — Total # with 1 null

```
In [23]:
#Map RegionDescription for Colorado Nation Parks to match the nomenclature of the national parks data set
geo_df["RegionDescription"]=geo_df["RegionDescription"].\
replace(["Great Sand Dunes National Park","Mesa Verde National Park","Black Canyon National Park","Rocky Mountai
        ["Great Sand Dunes NP & PRES","Mesa Verde NP","Black Canyon of the Gunnison NP","Rocky Mountain NP"])

In [24]:
#Check that mapped RegionDescription changes have been applied
geo_df["RegionDescription"].unique()

Out[24]: array(['Colorado National Monument', 'PAGOSA Ranger District (CO)',
       'SULPHUR RD - FS', 'Comanche RD-FS', 'Great Sand Dunes NP & PRES',
       'DILLON RD - FS', 'CLEAR CREEK RD -FS', 'SAGUACHE RD -FS',
       'CONEJOS PEAK RD -FS', 'LEADVILLE RD - FS', 'PIKES PEAK RD -FS',
       'Curecanti National Recreation Area', 'SOUTH PARK RD -FS',
       'Mesa Verde NP', 'ASPEN RD -FS', 'HAHNS P/BEARS EARS RD -FS',
       'SAN CARLOS RD -FS', 'NORWOOD RD - FS', 'CANYON LAKES RD -FS',
```

```
In [40]:
#Drop null values for FacilityID
geo_clean_df=geo_df.loc[geo_df['FacilityID'].notna()==True,:]
geo_clean_df

Out[40]:
```

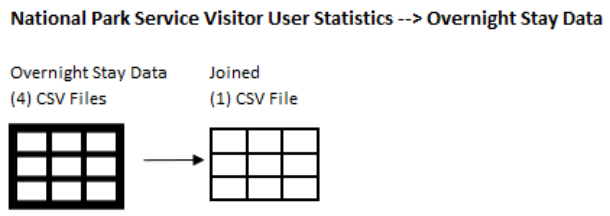
Updated dataframe

	RegionDescription	Park	FacilityID	FacilityState	FacilityLongitude	FacilityLatitude	CityPlace	County
0	Colorado National Monument	SADDLEHORN CAMPGROUND	234778.0	CO	-108.733	39.1048	Fruita	Mesa County
1	PAGOSA Ranger District (CO)	BLANCO RIVER GROUP CAMPGROUND	233318.0	CO	-106.883	37.1453	Pagosa Springs	Archuleta County
2	SULPHUR RD - FS	CUTTHROAT BAY GROUP	233880.0	CO	-105.874	40.1914	Granby	Grand County
...	...	...	...	...	...	...	...	...
194	GRAND VALLEY RD-FS	ASPEN LEAF	234664.0	CO	-108.674	38.6506	Orchard Mesa	Mesa County
195	SAN CARLOS RD -FS	DAVENPORT CAMPGROUND	234684.0	CO	-105.070	38.0556	Colorado City	Pueblo County
196	GRAND VALLEY RD-FS	BLACK BEAR	233216.0	CO	-108.081	39.0522	Cedaredge	Delta County

196 rows × 8 columns — Total # with removing null

One null value in Facility ID was detected in the Geocode data. This row of data was removed.

RegionDescription	Park	FacilityID	FacilityState	FacilityLongitude	FacilityLatitude	CityPlace	County
Colorado National M	SADDLEHORN CAM	234778	CO	-108.733	39.1048	Fruita	Mesa County
PAGOSA Ranger Dis	BLANCO RIVER GR	233318	CO	-106.883	37.1453	Pagosa Springs	Archuleta County
SULPHUR RD - FS	CUTTHROAT BAY G	233880	CO	-105.874	40.1914	Granby	Grand County
Comanche RD-FS	Picket Wire Canyonla	234166	CO	-103.524	37.9857	La Junta	Otero County
Great Sand Dunes N	Pinon Flats Campgrc	234685	CO	-105.517	37.7333	Westcliffe	Custer County

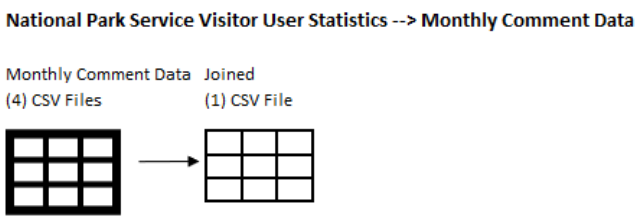


The transforming of the National Park Service Visitor User Statistics / Overnight Stay Data was done locally using Excel. The files were much more manageable each consisting of 133 rows and 40 columns of data.

Column headers were renamed to clarify descriptions (e.g. 'field 1' renamed to 'park'), a 'state' column was added to include 'CO', and an 'ID' to provide a primary key, ~25 static 'sum' columns were deleted. The four csv files were joined into one csv file with the following fields in our dataset: ID, Park, State, Year, Month, Recreation_Visitors, Non_Recreation_Visitors, Concession_Lodging, Tent_Campers, RV_Campers, Backcountry_Campers, Misc_Campers, Total_Overnight_Stays.

ID	Park	State	Year	Month	Recreation_	Non_Recreat	Recreation_V	Non_Recreat	Concession_I	Tent_Campe	RV_Campers	Backcountry_	Misc_Campe	Total_Overnight_Stays
1	Black Canyor	CO	2010	January	2,519	0	6,898	0	0	19	8	0	0	27
2	Black Canyor	CO	2010	February	3,034	0	9,193	0	0	35	13	0	0	48
3	Black Canyor	CO	2010	March	3,199	0	8,741	0	0	35	20	0	0	55
4	Black Canyor	CO	2010	April	4,421	0	17,747	0	0	169	160	32	0	361
5	Black Canyor	CO	2010	May	21,259	0	128,429	0	0	1,159	715	107	0	1,981
6	Black Canyor	CO	2010	June	24,477	0	163,138	0	0	1,172	1,518	184	0	2,874
7	Black Canyor	CO	2010	July	35,398	0	228,706	0	0	2,162	1,648	143	0	3,953
8	Black Canyor	CO	2010	August	27,367	0	165,614	0	0	1,095	1,293	196	0	2,584
9	Black Canyor	CO	2010	September	24,297	0	161,738	0	0	1,455	1,001	320	0	2,776
10	Black Canyor	CO	2010	October	21,331	0	75,893	0	0	803	630	159	0	1,592
11	Black Canyor	CO	2010	November	3,646	0	9,242	0	0	50	75	5	0	130
12	Black Canyor	CO	2010	December	5,396	0	10,577	0	0	11	8	0	0	19

Data above is for Black Canyon data only. Same structure for three other parks.

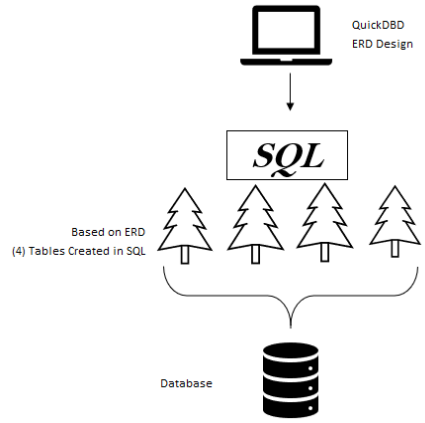


The transforming of the National Park Service Visitor User Statistics / Visitor Comments data was also done locally using Excel. Each file contained 20-98 rows and only 2 columns of data.

We added three columns 'ID', 'Park' and 'Year' to provide a primary key to join tables. Excel was then used to consolidate the four csv files into one file containing the following fields: ID, Park, State, Year, CollectedDate, and Comments.

ID	Park	State	Year	CollectedDat	Comments
1	Mesa Verde	CO	2010	1/1/10	Park closed January 21-24 due to weather.
2	Great Sand D	CO	2010	3/1/10	Our campground was closed in March 2010 for septic tank & sewer line construction.
3	Great Sand D	CO	2010	6/1/10	There was a forest fire in the northern part of the park in June this year. Mosca, Medano & Music Passes were closed for extended
4	Mesa Verde	CO	2010	10/1/10	Morefield Campground closed for the season Oct 13 and Far View Lodge closed October 21.
5	Rocky Mount	CO	2010	10/1/10	Much of October 2010 was unseasonably warm, especially compared to October 2009, when the last day Trail Ridge Road was ope
6	Mesa Verde	CO	2010	11/1/10	Wetherill Mesa, Far View Lodge, and Morefield Campground closed for winter season.
7	Mesa Verde	CO	2010	12/1/10	Winter closures for lodge, campground, and Wetherill Mesa.
8	Mesa Verde	CO	2011	1/1/11	Campground and lodge closed for winter.
9	Mesa Verde	CO	2011	4/1/11	Far View Lodge opened 4/21/11, Morefield CG and Wetherill Mesa still closed.
10	Black Canyor	CO	2011	4/1/11	Lots of rain and stormy weather, esp. on weekends.
11	Mesa Verde	CO	2011	5/1/11	Edited per conversation with Butch Street regarding skewed counter data. 9/21/2011

Load:

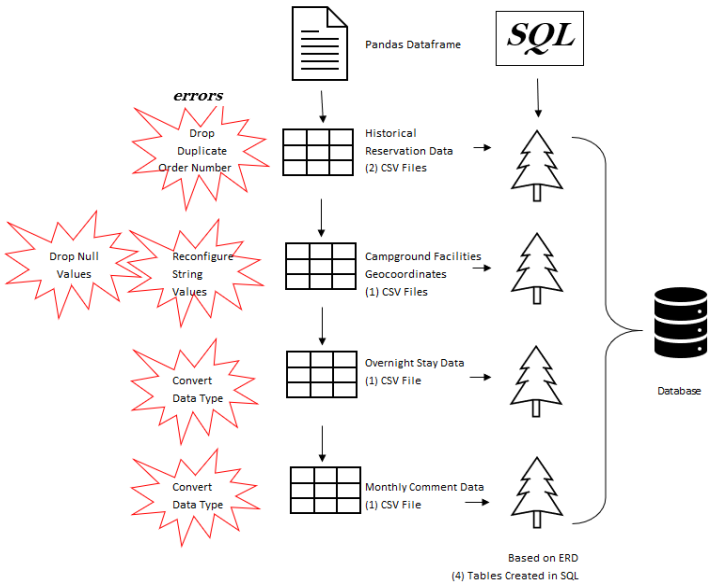


The database platform used to load the cleaned csv files was PostgreSQL. We determined that PostgreSQL would be the best database solution as the reservation data tables and the national parks data files had relational characteristics. An ERD (Entity Relationship Diagram) was developed for this database using Quick DBD and was imported into a SQL database called 'colorado_camping_db'. To load the data into this SQL database, we accessed a jupyter notebook where pandas and sqlalchemy

were utilized to upload the cleaned data from RIDB reservations and NPS files.

During the load process, troubleshooting issues arose, which caused additional transformation to occur. These issues included incorrect data types in NPS files, differences in National Park nomenclature

in the two data sources, and duplicates in RIDB data. For the NPS files, all but one datatype were string objects, even for columns with integer values. To remedy this before loading into our database, a dictionary was created using pandas to reassign desired data types.

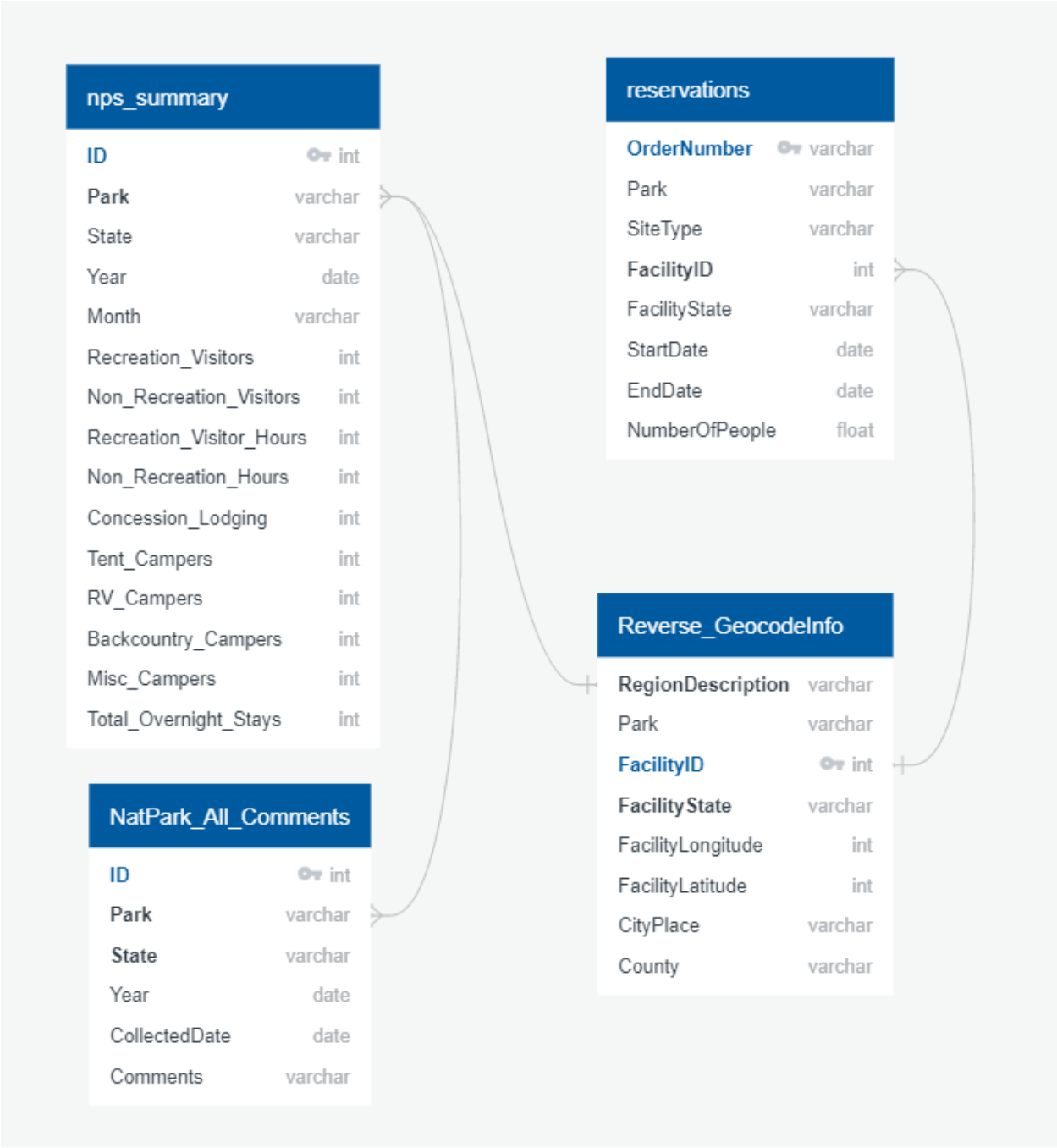


Upon further inspection of the proposed tables within our database, connecting RIDB coordinate data on the column 'RegionDescription' with the 'Park' column on NPS data was hindered by the value referral of National Parks being different in both datasets. This was corrected by filtering for the unique values in the columns for each dataset and differences were replaced with the same values per mapping.

Loading the two datasets continued, but upon further queries into the generated SQL database, duplicate values for the RIDB data was found in 'OrderNumber' for years between 2016-2020. This was determined to be a small anomaly of a repeated order number, which we effectively dropped from the dataframe within the jupyter notebook. Per this transformation, we had to drop the 'reservations' table from the database, upload it again into the SQL database, and confirm that duplicates were removed.

We discovered the transform and load process was iterative and required several cycles between transform and load.

Entity Relationship Diagram (ERD):



Optional Scenarios/Related Queries:

Below are examples of common questions users may have and how the database can be queried to provide answers

Scenarios & Related Queries	Tables source in SQL database	SQL Database Query
1. Kevin wants to go tent camping with 15 family members. Where can he camp in Colorado?	reservations	

```
-- Kevin wants to go tent camping with 15 family members. Where can he camp in Colorado?

SELECT * FROM reservations
WHERE "NumberOfPeople" >= 15 and "SiteType" = 'TENT ONLY NONELECTRIC';
```

Data Output | Explain | Messages | Notifications

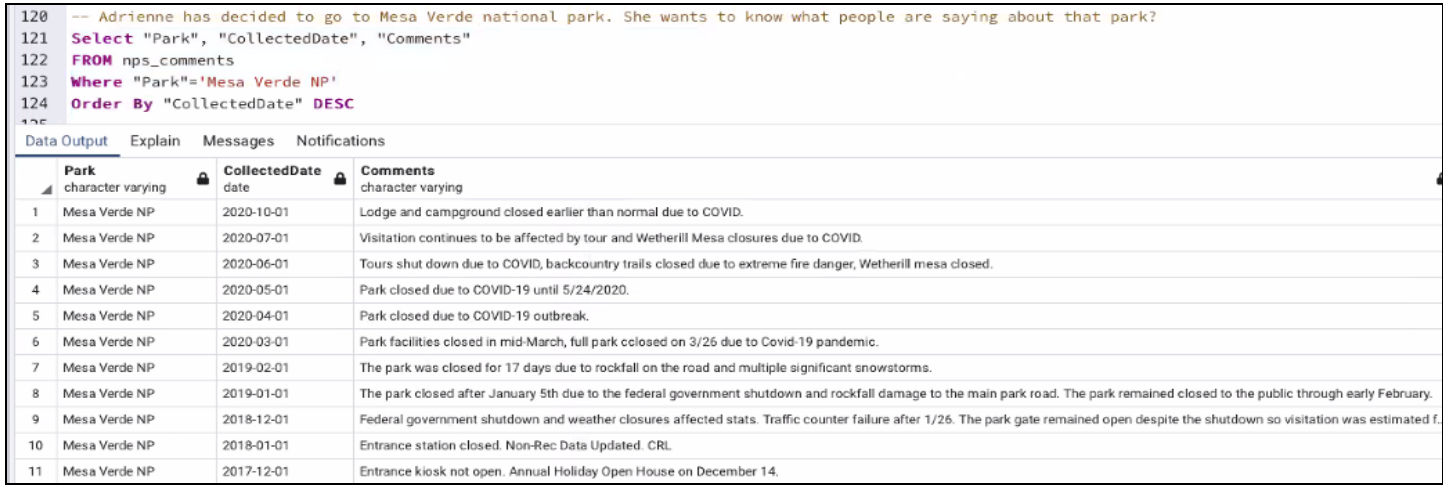
	OrderNumber [PK] character varying	Park character varying	SiteType character varying	FacilityID integer	FacilityState character varying	StartDate date	EndDate date	NumberOfPeople double precision
1	2-37121942	GREEN RIDGE	TENT ONLY NONELECT...	231861	CO	2017-06-30	2017-07-04	15
2	2-37341353	GREEN RIDGE	TENT ONLY NONELECT...	231861	CO	2017-06-16	2017-06-17	16
3	2-37580032	GREEN RIDGE	TENT ONLY NONELECT...	231861	CO	2017-08-04	2017-08-06	16
4	2-37724796	SOUTH MEADOWS	TENT ONLY NONELECT...	232339	CO	2017-06-09	2017-06-11	16
5	2-37745026	GREEN RIDGE	TENT ONLY NONELECT...	231861	CO	2017-08-04	2017-08-06	16
6	2-37832650	GREEN RIDGE	TENT ONLY NONELECT...	231861	CO	2017-07-22	2017-07-23	24
7	2-37892107	GREEN RIDGE	TENT ONLY NONELECT...	231861	CO	2017-08-10	2017-08-11	22
8	2-38019378	GREEN RIDGE	TENT ONLY NONELECT...	231861	CO	2017-06-26	2017-06-30	16
9	2-38443987	GREEN RIDGE	TENT ONLY NONELECT...	231861	CO	2017-08-11	2017-08-13	15
10	2-38958531	GREEN RIDGE	TENT ONLY NONELECT...	231861	CO	2017-08-26	2017-08-27	15

2. Sara loves tent camping and wants to share the experience with the entire class! She wants to know if there is a location at the Sand Dunes park where they can all tent camp (in separate tents of course).	reservations + geocodeInfo	

```
-- Sara loves tent camping, but only with a group of friends! She wants to know if there is a location at Sand Dunes where she can tent camp.
select "Park","FacilityID", "SiteType"
from reservations
where "SiteType"='GROUP TENT ONLY AREA NONELECTRIC'
AND "FacilityID" IN
    (select  "FacilityID"
     from geocode_info
     where  "RegionDescription"='Great Sand Dunes NP & PRES')
Group BY "Park", "FacilityID", "SiteType"
```

Data Output | Explain | Messages | Notifications

	Park character varying	FacilityID integer	SiteType character varying
1	Pinon Flats Campground	234685	GROUP TENT ONLY AREA NONELECTRIC

Scenarios & Related Queries	Tables source in SQL database	SQL Database Query
3. Leah hates crowds but loves camping. She wants to find out which campground (park) is the least busy at Rocky Mountain National Park.	reservations	
4. Adrienne has decided to go to Mesa Verde national park. She wants to know what people are saying about that park.	NatPark_All_Comments	

```sql
-- Kevin hates crowds but loves camping. He wants to find out which campground is the least busy at rocky mountian national park. Rocky Mountain NP
select "Park","FacilityID", count("OrderNumber")
from reservations
where "FacilityID" IN
    (select  "FacilityID"
     from geocode_info
     where  "RegionDescription"='Rocky Mountain NP')
Group BY "Park", "FacilityID"
Order BY count("OrderNumber") ASC
```

Data Output | Explain | Messages | Notifications

	Park character varying	FacilityID integer	count bigint
1	Aspenglen Campground	233187	3966
2	Glacier Basin (CO)	232462	8855
3	Moraine Park (CO)	232463	18309
4	ASPENGLEN CAMPGROUND	233187	35714
5	GLACIER BASIN CAMPGROUND	232462	57120
6	MORAINE PARK CAMPGROUND	232463	149957

```sql
-- Adrienne has decided to go to Mesa Verde national park. She wants to know what people are saying about that park?
Select "Park", "CollectedDate", "Comments"
FROM nps_comments
Where "Park"='Mesa Verde NP'
Order By "CollectedDate" DESC
```

Data Output | Explain | Messages | Notifications

	Park character varying	CollectedDate date	Comments character varying
1	Mesa Verde NP	2020-10-01	Lodge and campground closed earlier than normal due to COVID.
2	Mesa Verde NP	2020-07-01	Visitation continues to be affected by tour and Wetherill Mesa closures due to COVID.
3	Mesa Verde NP	2020-06-01	Tours shut down due to COVID, backcountry trails closed due to extreme fire danger, Wetherill mesa closed.
4	Mesa Verde NP	2020-05-01	Park closed due to COVID-19 until 5/24/2020.
5	Mesa Verde NP	2020-04-01	Park closed due to COVID-19 outbreak.
6	Mesa Verde NP	2020-03-01	Park facilities closed in mid-March, full park cclosed on 3/26 due to Covid-19 pandemic.
7	Mesa Verde NP	2019-02-01	The park was closed for 17 days due to rockfall on the road and multiple significant snowstorms.
8	Mesa Verde NP	2019-01-01	The park closed after January 5th due to the federal government shutdown and rockfall damage to the main park road. The park remained closed to the public through early February.
9	Mesa Verde NP	2018-12-01	Federal government shutdown and weather closures affected stats. Traffic counter failure after 1/26. The park gate remained open despite the shutdown so visitation was estimated f...
10	Mesa Verde NP	2018-01-01	Entrance station closed. Non-Rec Data Updated. CRL
11	Mesa Verde NP	2017-12-01	Entrance kiosk not open. Annual Holiday Open House on December 14.