DATA 2001 ASSIGNMENT

1. Dataset description

The original five files are all from Canvas. (https://canvas.sydney.edu.au/courses/30951/modules)

Neighbourhoods: There are nine columns in this dataset: index, area_id, area_name, land_area, population, number_of_dwellings, number_of_businesses median_annual_household_income and average_monthly_rent. The index presents the number of each row in this dataset, and it was added automatically by the server. Both area-id and area name are which area these data represent. The other columns present the related data in this area. For example, the population represents how many people are in this area.

StatisticalAreas: There are three columns in this file, which are area_id, parent_id and area name. We found some duplicate values in this file, so we use the drop_duplicates() function to clean it.

BusinessStats: There are eight columns here, which are area-id, number of businesses, accommodation and food, retail trade, agriculture forestry and fishing, health care and social assistance, public administration and safety, transport postal and warehousing. These data represent the income and economic level in such neighbourhoods.

As for two shp files, we use the PostGIS to upload them. These two files represent some geography data. By using PostGIS to upload the shp file, it automatically adds the gid in the table, which is like the spatial index.

Rainfall: We use the following code to select the area which is all in these five tables. As Australian Government Bureau of Meteorology (n.d.) shows there are many data about rainfall for many stations. And we find the nearest station with more than 20 years of data and are still open or closed after 2010. Then we record the average rainfall for all months and annual. Some areas have few stations far away from the neighbour or have little data that couldn’t represent the average rainfall, so we regard them as NaN values in the file.

There are 15 columns in the rainfall csv, which are 12 months, annual, area_id and area_name. The month column means the rainfall in that month; for example, the data in Jan represents the rainfall in this area in January.

Data cleaning:

Then, we know that the population, shape_area, rainfall, average rent, ASSISTANCE and income couldn' t be negative, so we add the constraints in selecting part. For exapmle,

```
#rainfall_density

rainfall_density_query = """
SELECT area_id, (R.annual) / N.land_area AS "rainfall_density"
    FROM rainfall R JOIN neighbourhoods N USING (area_id)
   WHERE R.annual IS NOT NULL
   AND R.annual >0;
"""


rainfall_density_pd = pd.read_sql_query(rainfall_density_query, conn)
print(rainfall_density_pd)
```

We directly upload the dataset and clean it in the calculation step. We used the drop_duplicates() function to clean the duplicated values, and we cleaned the nan value by dropna() function.

2. Database description

There are seven datasets in the database: Neighbourhoods, SA2_2016_AUST, BusinessStats, StatisticalAreas, RFSNSW_BFPL, Rainfall and final table. The primary key for the Neighbourhoods, BusinessStats, StatisticalAreas, and Rainfall is area_id and these tables connected by it. However, we found that in statisticalarea table, there are several duplicated values both in area_ia and other columns, so we use the drop_duplicated() function to drop them to make sure the primary key is unique. So the foreign key is also area_id in these tables. As for the SA2_2016_AUST and RFSNSW_BFPL, the primary key is gid, and they also connect by it, so its foreign key is also it. To connect Neighbourhoods with SA2_2016_AUST, we find sa2_name16 is the same as the area_name, so we set the area_name as its foreign key.
We use the pgadmin to set the primary key and the foreign key. As the primary ket must be unique, we first add the constraints to make sure all area_id and gid are unique, the screenshots are in the appendix. Then we set the PK in primary key page from Constraints. After setting all primary keys, we select the columns which we will use to connect with other tables as the fo

reign key in the foreign key page from constraints. All the steps are in the appendix.

The ERD Diagram is in appendix.

3. Bush fire risk analysis

We calculate the population density by dividing the population by the land area. Then the calculation method is the same for the business density, dwellings density, and rainfall density. As for the assistive service density, we first add the values from health_care_and_social_assistance and public_administration_and_safety and use these results to divide by the land area, and we regard this as its assistive service density.

As for the bfpl density, as two factors may influence the density. So under this situation, we multiplicate these two factors, and then we use this result to divide the land area and regard these as the bfpl density.

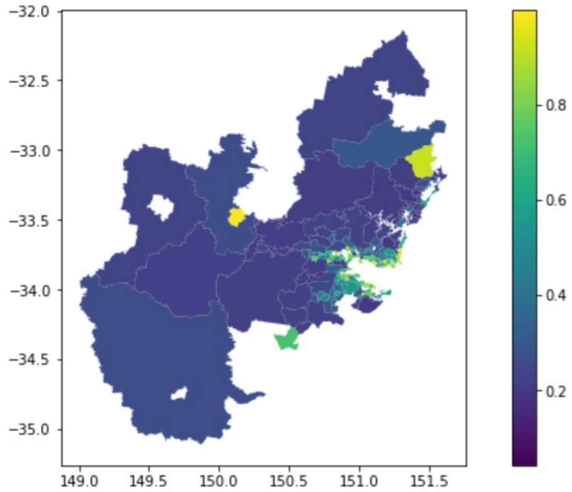
Then we make all these density columns into a new data frame called combina_density. If there are some nan or zero values, we use the dropna function to clean it before further analysis. What's more, we also drop the duplicated area_id by using the drop_duplicates() part.
Then we calculate its average value and the standard deviation for each density. Finally, we create new columns called the z_xx_density and insert the values calculated using the initial density minus the mean and divided by the standard deviation. We regard it as the final density for all different densities.

Then we use the following formula to calculate the z-score and the fire risk score. We use the
combina_density['z_population_density'] + combina_density['z_dwellings_density'] + combina_density['z_business_density'] + combina_density['z_bfpl_density'] - combina_density['z_assistive_service_density']-combina_density['z_rainfall_density']
formula to calculate it, and we find that most numbers are negative, which may cause misunderstanding so we use 0 to minus this result to make the most of these results positive.

As we should use the 1/1+e^(-t) to calculate the z score. So we should use the 0 minus z score and regard it as -t. Then we named this negative z score as total_negative_z_score. Then we regard the results of

1 / (1 + np.exp(combina_density['total_negative_z_score']))
as the fire risk.
We use the mean() function to calculate the average fire score, and we found it is about 0.4678, which is less than 0.5. So, we could say that the fire risk in NSW is low.

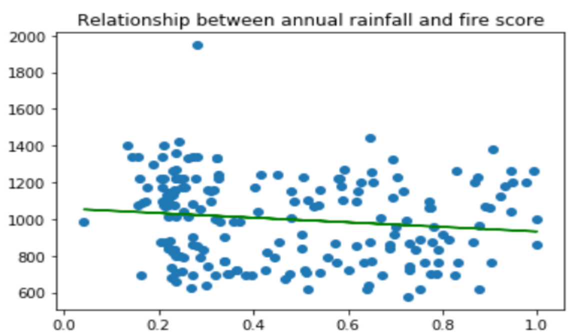
Then we coloured the neighbours in the NSW map by using fire score.



4. Correlation analysis

For the rainfall:

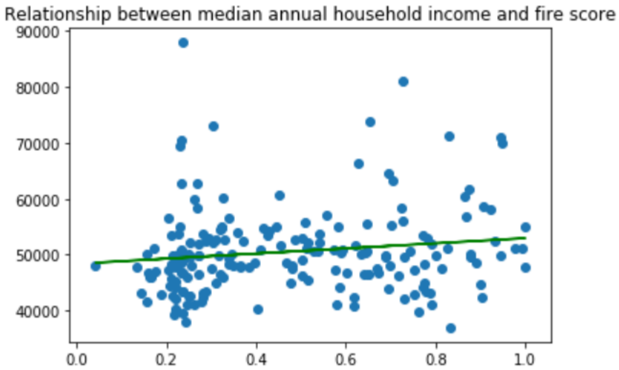
Volatility in conditions, from rains to extreme drought, can prime the land for fire. Meanwhile, in western Montana and other conifer forests in the Northwest, the opposite pattern occurred. CLIMATECENTRAL (2011) has explained that there is no relation between the fire possibility and the annual rainfall. The correlation between the fire score and yearly rainfall is about 0.00966. According to the scatter plot, we could find that the spray is almost random, proving that there is practically no relation between annual rainfall and fire possibility.



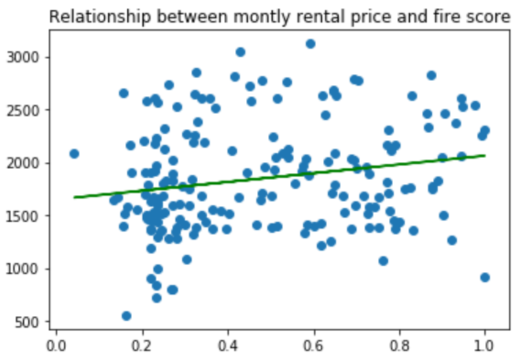
Income and rental price:

Low-income neighborhoods, as according to Schulz & Williams & Israel & Lempert (2002) demonstrate that tend to have a declining tax base, which results in the deterioration of public safety systems including police, firefighting, and the enforcement of regulations against illegal dumping. Krieger & Higgins (2002) revealed that houses in such areas often have hazardous cooking facilities and a lack of storage space, leading to clutter that can contribute to fire.

We knew that the correlation between income and the fire risk is about 0.16 by .corr() function. According to the scatter plot, we could see that though there are lots of points out of the fit line, the tendency of the scatter and the bar is almost the same, so we could say that there is a weak relationship between income and fire risk.



We knew that the correlation between the rental price and the fire risk is about 0.296. We could see a fitted line according to the plot, with the function is y=1646.71 + 413.37x. We found that the most scatter' s tendency is as same as the fit line, so, we could say that the lower the rental price, the higher the fire risk is.

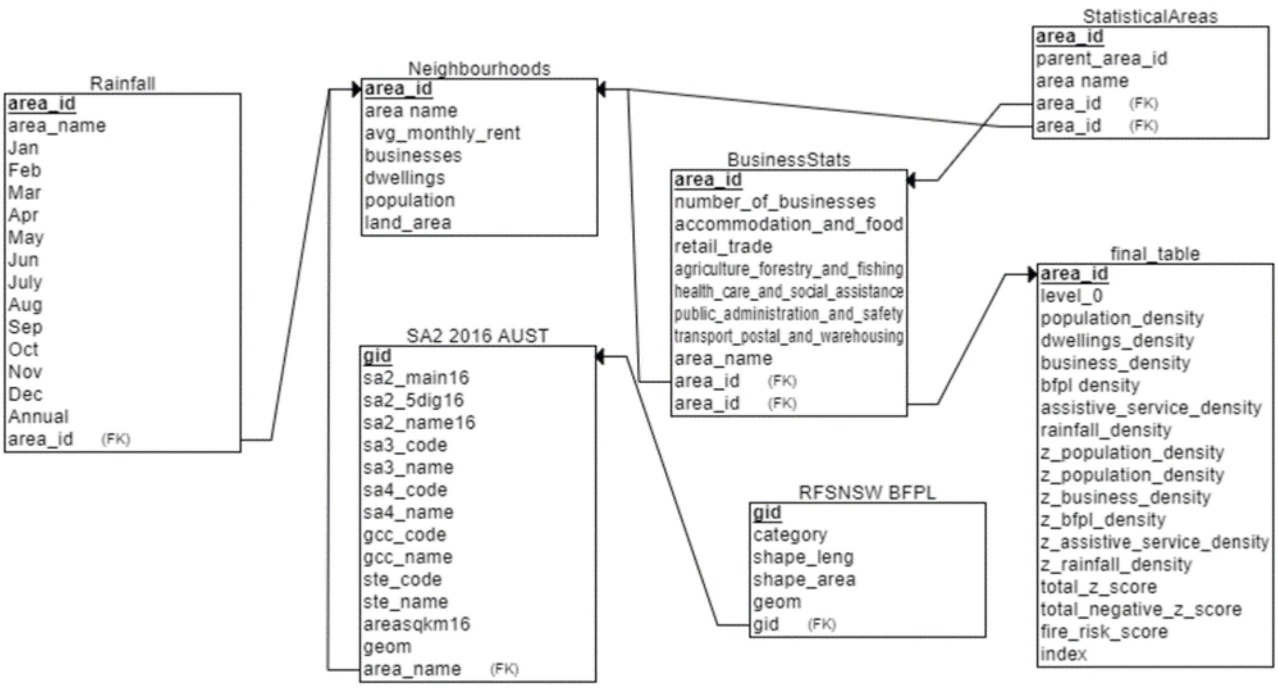


References:

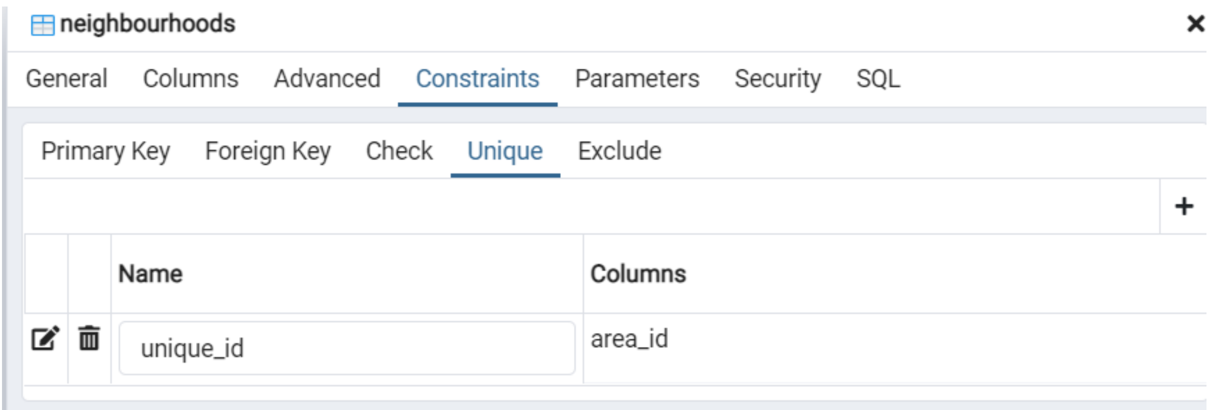
1. Australian Government Bureau of Meteorology. (n.d.). *Climate Data Online*. Climate Data Online. Retrieved May 5, 2021, from http://www.bom.gov.au/climate/data/index.shtml?bookmark=136
2. CLIMATECENTRAL. (2011, April 27). *Can Rain Cause More Fire?* https://www.climatecentral.org/gallery/graphics/can-rain-cause-more-fire
3. Schulz, A. J., Williams, D. R., Israel, B. A., & Lempert, L. B. (2002). Racial and spatial relations as fundamental determinants of health in Detroit. The Milbank Quarterly, 80(4), 677–iv. https://doi.org/10.1111/1468-0009.00028
4. Krieger, J., & Higgins, D. L. (2002). Housing and health: time again for public health action. American journal of public health, 92(5), 758–768. https://doi.org/10.2105/ajph.92.5.758

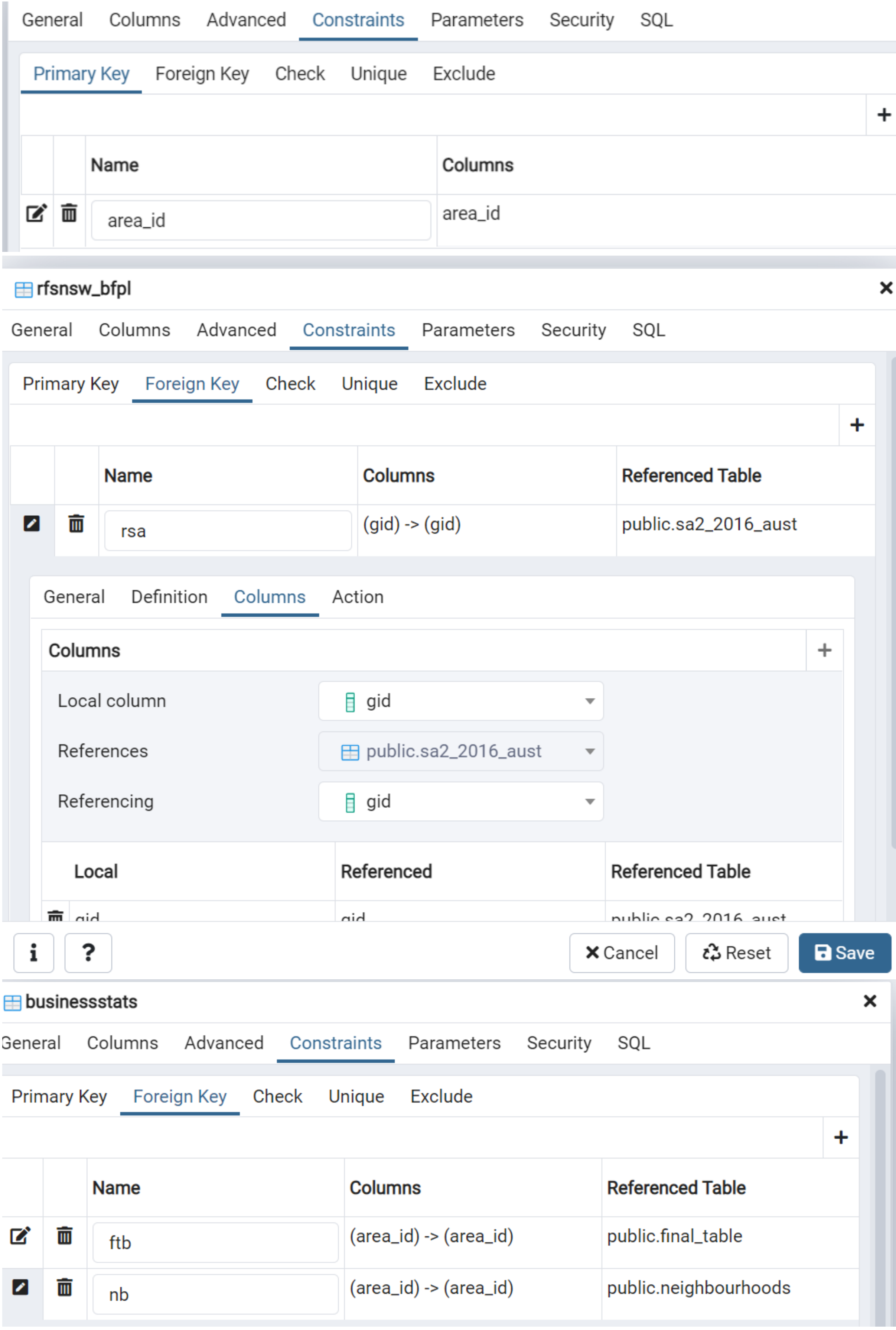
Appendix:

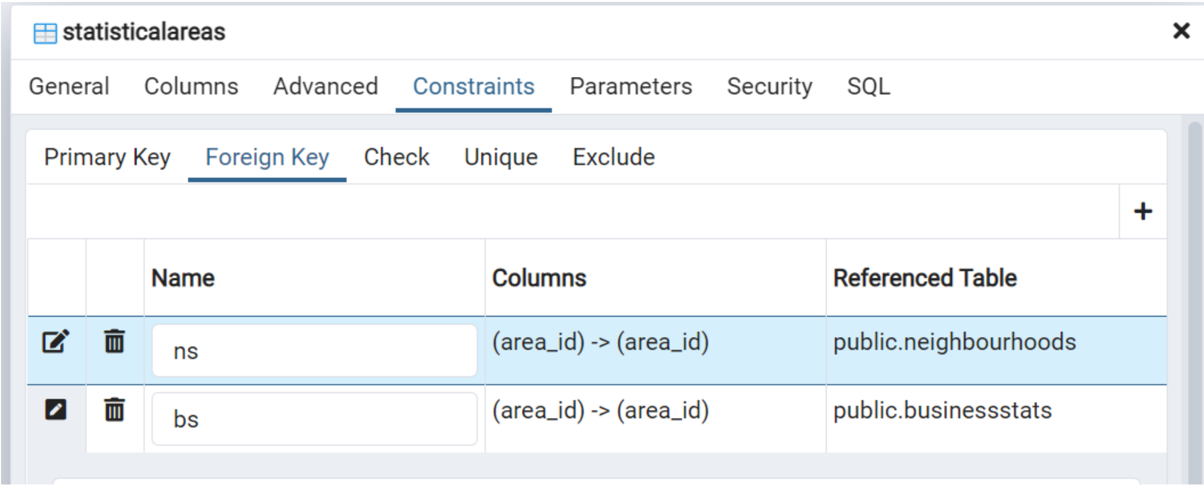
ERD:

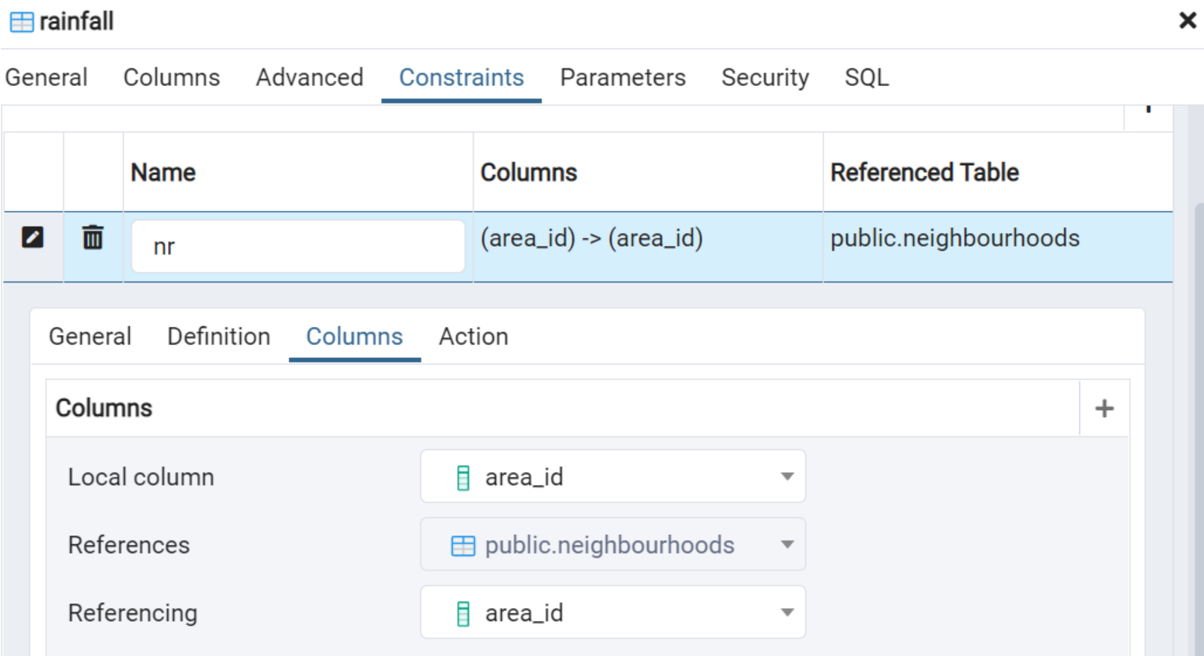


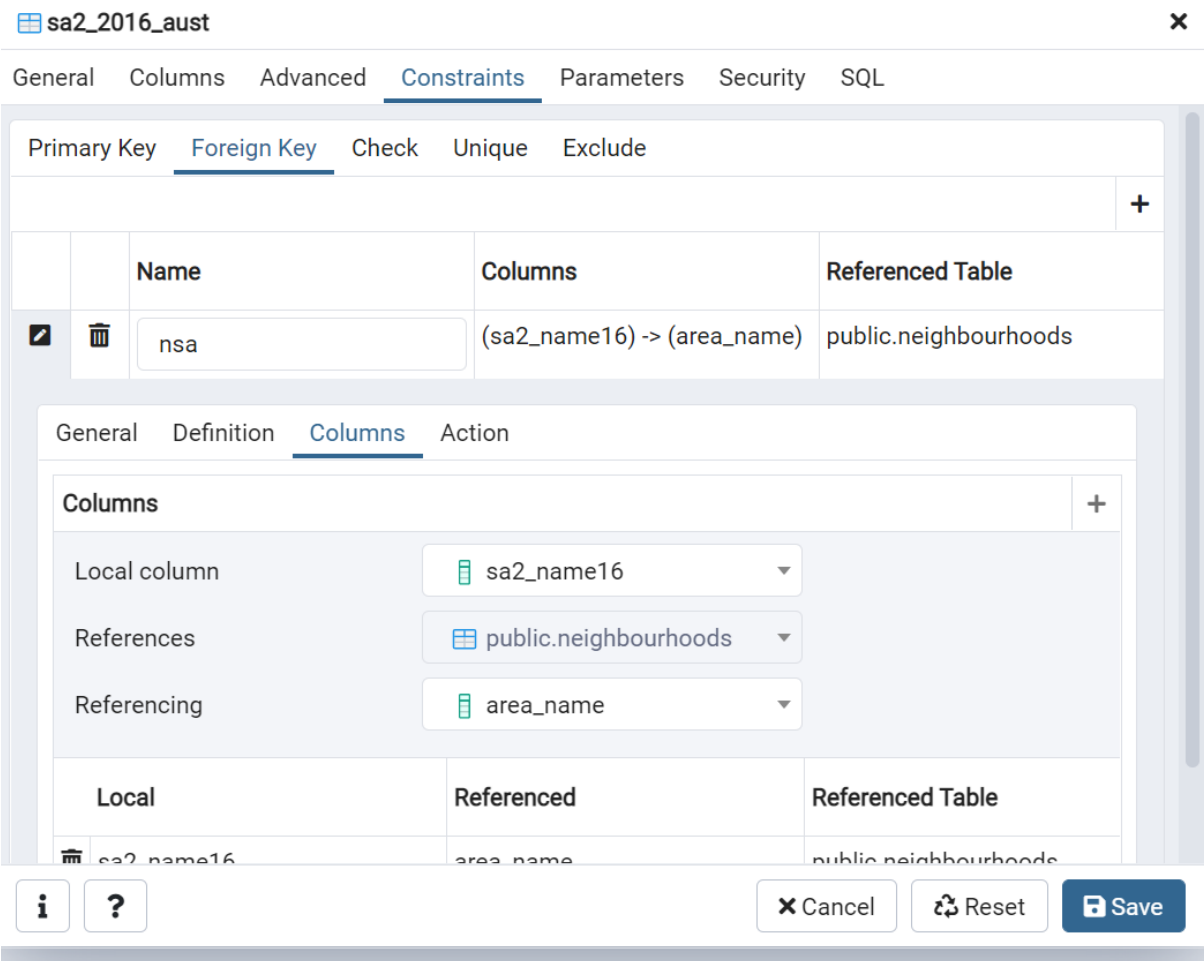
How to set the unique, primary key and foreign key:




General Columns Advanced Constraints Parameters Security SQL
Primary Key Foreign Key Check Unique Exclude
Name Columns
area_id area_id
rfsnsw_bfpl
General Columns Advanced Constraints Parameters Security SQL
Primary Key Foreign Key Check Unique Exclude
Name Columns Referenced Table
rsa (gid) -> (gid) public.sa2_2016_aust
General Definition Columns Action
Columns
Local column gid
References public.sa2_2016_aust
Referencing gid
Local Referenced Referenced Table
Cancel Reset Save
businessstats
General Columns Advanced Constraints Parameters Security SQL
Primary Key Foreign Key Check Unique Exclude
Name Columns Referenced Table
ftb (area_id) -> (area_id) public.final_table
nb (area_id) -> (area_id) public.neighbourhoods


statisticalareas
General
Columns
Advanced
Constraints
Parameters
Security
SQL
Primary Key
Foreign Key
Check
Unique
Exclude
Name
Columns
Referenced Table
ns
(area_id) -> (area_id)
public.neighbourhoods
bs
(area_id) -> (area_id)
public.businessstats


rainfall
General
Columns
Advanced
Constraints
Parameters
Security
SQL
Name
Columns
Referenced Table
nr
(area_id) -> (area_id)
public.neighbourhoods
General
Definition
Columns
Action
Columns
Local column
area_id
References
public.neighbourhoods
Referencing
area_id


sa2_2016_aust
General
Columns
Advanced
Constraints
Parameters
Security
SQL
Primary Key
Foreign Key
Check
Unique
Exclude
Name
Columns
Referenced Table
nsa
(sa2_name16) -> (area_name)
public.neighbourhoods
General
Definition
Columns
Action
Columns
Local column
sa2_name16
References
public.neighbourhoods
Referencing
area_name
Local
Referenced
Referenced Table
Cancel
Reset
Save