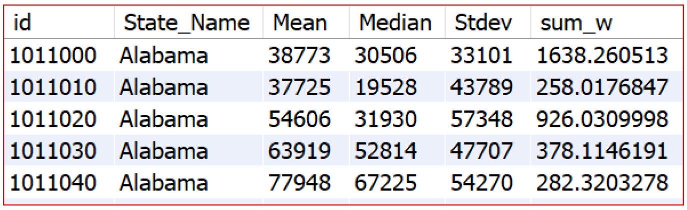
Data Cleaning and Exploratory Data Analysis for "US Household Income" dataset.

Project Overview

This project presents an exploratory data analysis (EDA) of US household income at the state, county, and city levels using SQL. The goal is to uncover geographic and economic insights by analyzing land/water distribution, income levels, and area classifications across different states and regions.

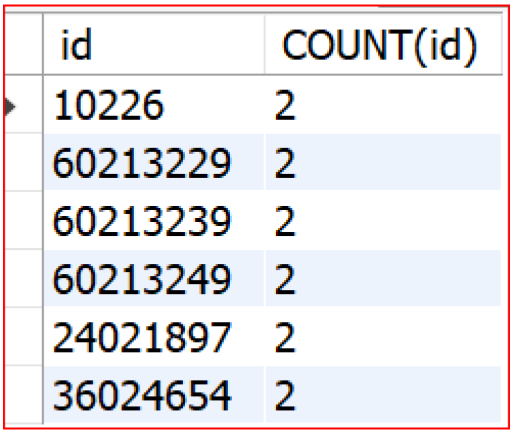
Data Cleaning in SQL:

```
-- The first column name need to update as 'id'
ALTER TABLE ushousehold_income_statistics RENAME COLUMN `ï»¿id` TO `id`;
```



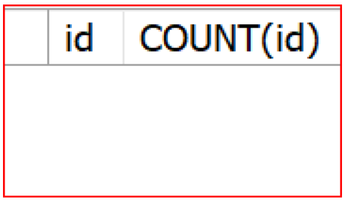
id	State_Name	Mean	Median	Stdev	sum_w
1011000	Alabama	38773	30506	33101	1638.260513
1011010	Alabama	37725	19528	43789	258.0176847
1011020	Alabama	54606	31930	57348	926.0309998
1011030	Alabama	63919	52814	47707	378.1146191
1011040	Alabama	77948	67225	54270	282.3203278

```
-- Count the duplicate id's
SELECT id, COUNT(id)
FROM us_household_income
GROUP BY id
HAVING COUNT(id) > 1;
```



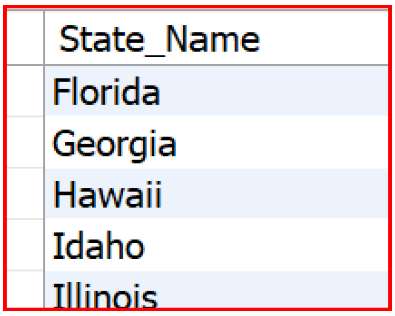
id	COUNT(id)
10226	2
60213229	2
60213239	2
60213249	2
24021897	2
36024654	2

```
-- Delete the duplicate records from the table --
DELETE FROM us_household_income
WHERE row_id IN (
SELECT row_id
FROM
(
SELECT row_id, id,
ROW_NUMBER() OVER(PARTITION BY id ORDER BY id) AS row_num
FROM us_household_income
) AS duplicates
WHERE row_num > 1);
```



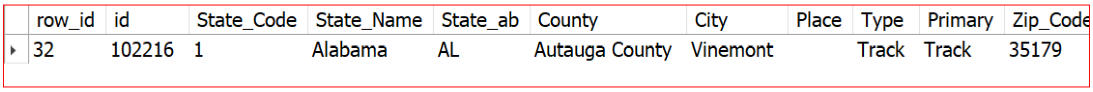
id	COUNT(id)

```
-- Update the state name
UPDATE us_household_income
SET State_Name = 'Georgia'
WHERE State_Name = 'georia';
```



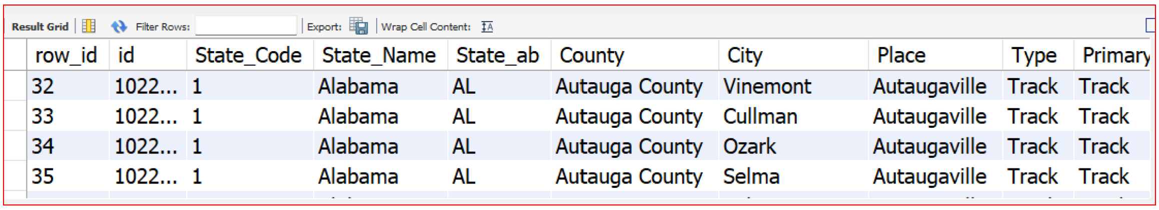
State_Name
Florida
Georgia
Hawaii
Idaho
Illinois

```
-- Check the blank value in Place column
SELECT *
FROM us_household_income
WHERE Place = ''
ORDER BY 1;
```



row_id	id	State_Code	State_Name	State_ab	County	City	Place	Type	Primary	Zip_Code
32	102216	1	Alabama	AL	Autauga County	Vinemont		Track	Track	35179

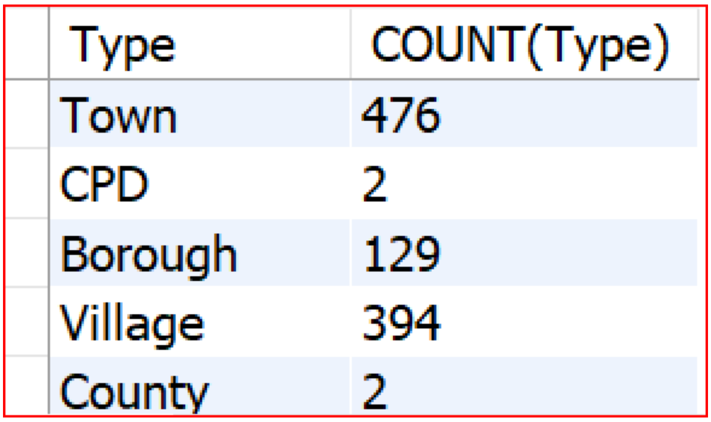
```
-- Update the db with actual value.
UPDATE us_household_income
SET Place = 'Autaugaville'
WHERE County = 'Autauga County'
AND City = 'Vinemont';
```



row_id	id	State_Code	State_Name	State_ab	County	City	Place	Type	Primary
32	1022...	1	Alabama	AL	Autauga County	Vinemont	Autaugaville	Track	Track
33	1022...	1	Alabama	AL	Autauga County	Cullman	Autaugaville	Track	Track
34	1022...	1	Alabama	AL	Autauga County	Ozark	Autaugaville	Track	Track
35	1022...	1	Alabama	AL	Autauga County	Selma	Autaugaville	Track	Track

```
-- Check the duplicate in Type column and update the table
SELECT Type, COUNT(Type)
FROM us_household_income
GROUP BY Type;

UPDATE us_household_income
SET Type = 'Borough'
WHERE Type = 'Boroughs';
```



Type	COUNT(Type)
Town	476
CPD	2
Borough	129
Village	394
County	2

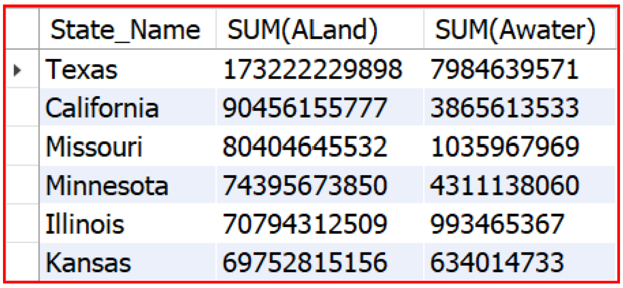
Exploratory Data Analysis:

```
-- Check the available records and structure in the us_household_income
dataset --
SELECT * FROM us_household_income;
```

```
-- Check the available records and structure in the
ushousehold_income_statistics dataset --
SELECT * FROM ushousehold_income_statistics;
```

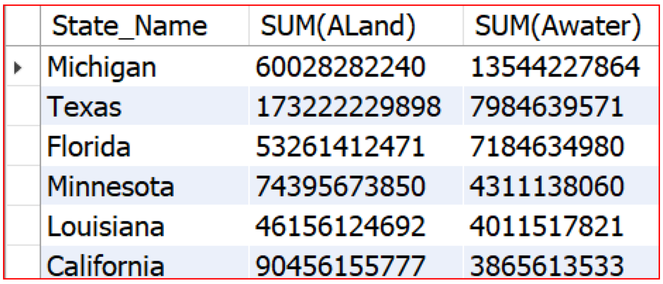
```
-- What are the land and water areas for each city and county within the
states --
SELECT State_Name,County, City, ALand, Awater
FROM us_household_income;
```

```
-- Which states have the largest total land area --
SELECT State_Name, SUM(ALand), SUM(Awater)
FROM us_household_income
GROUP BY State_Name
ORDER BY SUM(ALand) DESC;
```



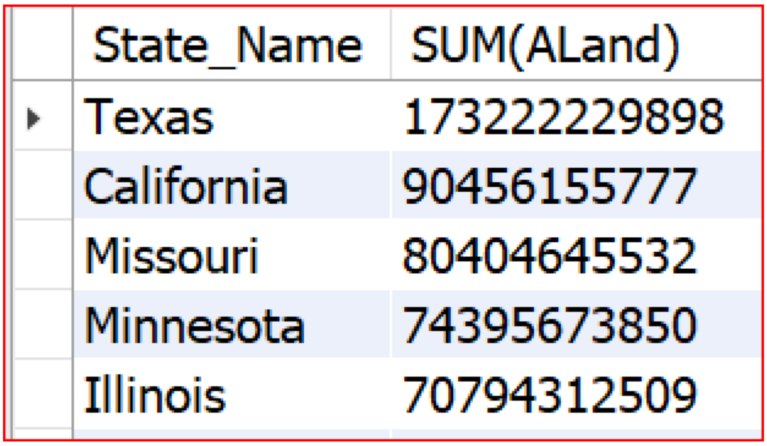
State_Name	SUM(ALand)	SUM(Awater)
Texas	173222229898	7984639571
California	90456155777	3865613533
Missouri	80404645532	1035967969
Minnesota	74395673850	4311138060
Illinois	70794312509	993465367
Kansas	69752815156	634014733

```
-- Which states have the largest total water area --
SELECT State_Name, SUM(ALand), SUM(Awater)
FROM us_household_income
GROUP BY State_Name
ORDER BY SUM(Awater) DESC;
```



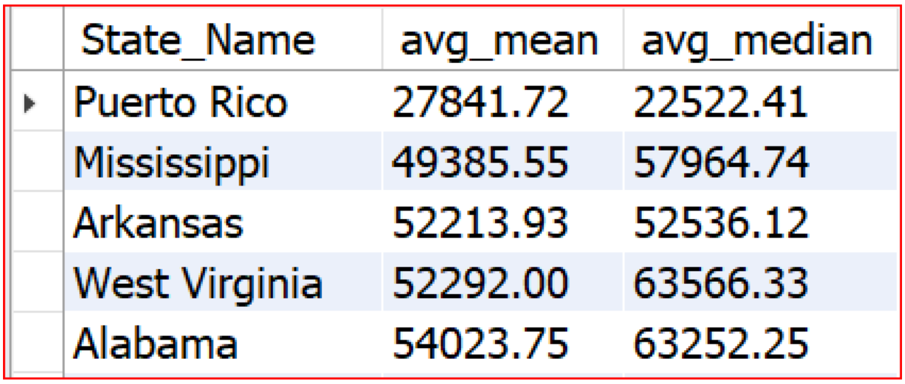
State_Name	SUM(ALand)	SUM(Awater)
Michigan	60028282240	13544227864
Texas	173222229898	7984639571
Florida	53261412471	7184634980
Minnesota	74395673850	4311138060
Louisiana	46156124692	4011517821
California	90456155777	3865613533

```
-- Which are the top 10 US states with the largest total land area --
SELECT State_Name, SUM(ALand)
FROM us_household_income
GROUP BY State_Name
ORDER BY SUM(ALand) DESC
LIMIT 10;
```



State_Name	SUM(ALand)
Texas	173222229898
California	90456155777
Missouri	80404645532
Minnesota	74395673850
Illinois	70794312509

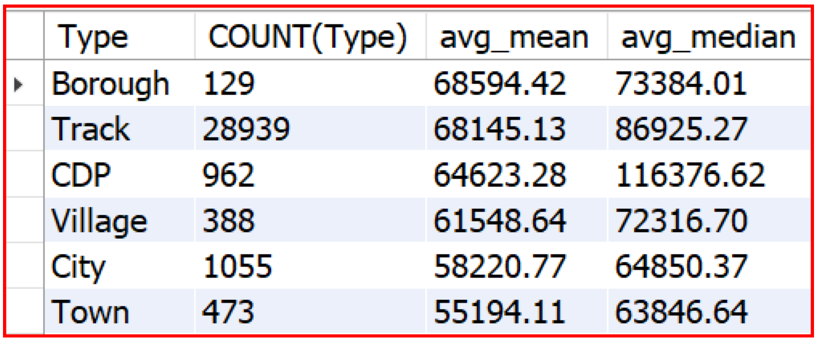
```
-- What is the average household income (mean and median) for each state --
SELECT ui.State_Name, ROUND(AVG(Mean), 2) AS avg_mean, ROUND(AVG(Median), 2)
AS avg_median
FROM us_household_income AS ui
INNER JOIN ushousehold_income_statistics AS us
ON ui.id = us.id
WHERE Mean <> 0
GROUP BY ui.State_Name
ORDER BY avg_mean;
```



State_Name	avg_mean	avg_median
Puerto Rico	27841.72	22522.41
Mississippi	49385.55	57964.74
Arkansas	52213.93	52536.12
West Virginia	52292.00	63566.33
Alabama	54023.75	63252.25

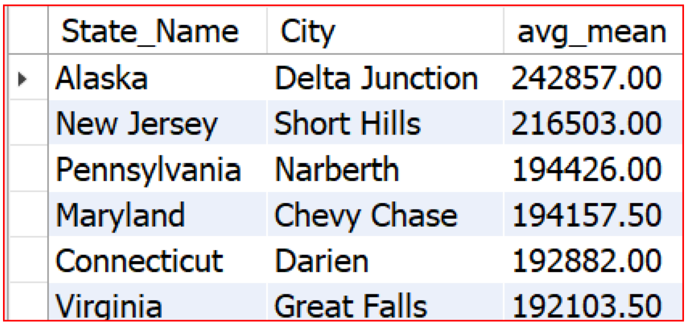
```
/* How does the average household income (mean and median) vary by area type
(e.g., urban, rural), considering only types with more than 100 records */

SELECT Type, COUNT(Type), ROUND(AVG(Mean), 2) AS avg_mean,
ROUND(AVG(Median), 2) AS avg_median
FROM us_household_income AS ui
INNER JOIN ushousehold_income_statistics AS us
ON ui.id = us.id
WHERE Mean <> 0
GROUP BY Type
HAVING COUNT(Type) > 100
ORDER BY avg_mean DESC;
```



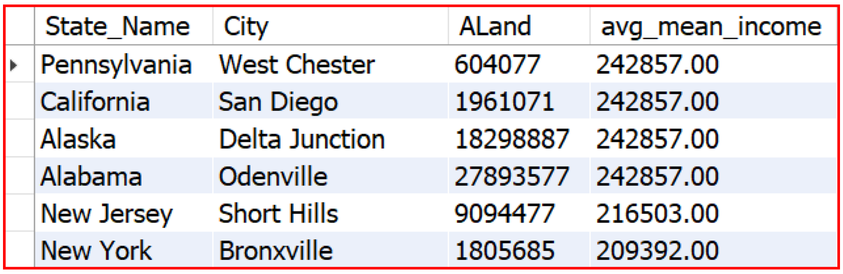
Type	COUNT(Type)	avg_mean	avg_median
Borough	129	68594.42	73384.01
Track	28939	68145.13	86925.27
CDP	962	64623.28	116376.62
Village	388	61548.64	72316.70
City	1055	58220.77	64850.37
Town	473	55194.11	63846.64

```
-- Which cities in the US have the highest average household income --
SELECT ui.State_Name, City, ROUND(AVG(Mean), 2) AS avg_mean
FROM us_household_income AS ui
INNER JOIN ushousehold_income_statistics AS us
ON ui.id = us.id
GROUP BY ui.State_Name, City
ORDER BY avg_mean DESC;
```



State_Name	City	avg_mean
Alaska	Delta Junction	242857.00
New Jersey	Short Hills	216503.00
Pennsylvania	Narberth	194426.00
Maryland	Chevy Chase	194157.50
Connecticut	Darien	192882.00
Virginia	Great Falls	192103.50

```
-- City Size vs Income --
SELECT ui.State_Name, City, ALand, ROUND(AVG(Mean), 2) AS avg_mean_income
FROM us_household_income AS ui
JOIN ushousehold_income_statistics AS us ON ui.id = us.id
GROUP BY ui.State_Name, City, ALand
ORDER BY avg_mean_income DESC;
```



State_Name	City	ALand	avg_mean_income
Pennsylvania	West Chester	604077	242857.00
California	San Diego	1961071	242857.00
Alaska	Delta Junction	18298887	242857.00
Alabama	Odenville	27893577	242857.00
New Jersey	Short Hills	9094477	216503.00
New York	Bronxville	1805685	209392.00