**SQL PROJECT DATA ANALYSIS**

**SQL Data Analysis Project: California Housing Dataset**

**Project Overview**

In this project, I performed **exploratory data analysis (EDA)** on the **California Housing dataset** using **SQL Server**. The dataset contains information about housing attributes across different regions of California, such as median income, population, number of rooms, proximity to the ocean, and median house values.

The objective was to gain actionable insights and uncover regional trends that could inform decision-making in real estate, urban planning, or housing affordability studies.

```sql
--1. Average Median House Value by Ocean Proximity
SELECT
ocean_proximity,
AVG(CAST (median_house_value AS BIGINT)) AS 'Average Median Value'
FROM
housing
GROUP BY ocean_proximity;
```

150 %

Results | Messages

| | ocean_proximity | Average Median Value |
|---|---|---|
| 1 | INLAND | 124805 |
| 2 | NEAR BAY | 259212 |
| 3 | NEAR OCEAN | 249433 |
| 4 | <1H OCEAN | 240084 |
| 5 | ISLAND | 380440 |

```sql
-- Distinct ocean proximity
SELECT DISTINCT ocean_proximity
FROM
housing;
```

150 %

Results | Messages

| | ocean_proximity |
|---|---|
| 1 | INLAND |
| 2 | NEAR BAY |
| 3 | NEAR OCEAN |
| 4 | <1H OCEAN |
| 5 | ISLAND |

```sql
--2. Top 5 Most Densely Populated Areas
SELECT TOP 5
longitude,
latitude,
population
FROM
housing
ORDER BY population DESC;
```

150 %

Results | Messages

| | longitude | latitude | population |
|---|---|---|---|
| 1 | -117.419998168945 | 33.3499984741211 | 35682 |
| 2 | -121.790000915527 | 36.6399993896484 | 28566 |
| 3 | -121.440002441406 | 38.4300003051758 | 16305 |
| 4 | -117.73999786377 | 33.8899993896484 | 16122 |
| 5 | -117.779998779297 | 34.0299987792969 | 15507 |

```sql
--3. Average Income vs. Average House Value by Region
SELECT
ocean_proximity,
AVG(CAST(median_income AS BIGINT)) AS 'Average income',
AVG(CAST(median_house_value AS BIGINT)) AS 'Average house value'
FROM
housing
GROUP BY ocean_proximity
ORDER BY AVG(CAST(median_income AS BIGINT)) DESC
```

150 %

Results | Messages

| | ocean_proximity | Average income | Average house value |
|---|---|---|---|
| 1 | NEAR BAY | 3 | 259212 |
| 2 | NEAR OCEAN | 3 | 249433 |
| 3 | <1H OCEAN | 3 | 240084 |
| 4 | ISLAND | 2 | 380440 |
| 5 | INLAND | 2 | 124805 |

```sql
--4. Correlation Proxy Between Income and House Value
SELECT
    (
        COUNT(*) * SUM(CAST(median_income AS FLOAT) * CAST(median_house_value AS FLOAT)) -
        SUM(CAST(median_income AS FLOAT)) * SUM(CAST(median_house_value AS FLOAT))
    ) /
    SQRT(
        (COUNT(*) * SUM(CAST(median_income AS FLOAT) * CAST(median_income AS FLOAT)) - POWER(SUM(CAST(median_income AS FLOAT)), 2)) *
        (COUNT(*) * SUM(CAST(median_house_value AS FLOAT) * CAST(median_house_value AS FLOAT)) - POWER(SUM(CAST(median_house_value AS FLOAT)), 2))
    ) AS PearsonCorrelation
FROM housing
WHERE median_income IS NOT NULL AND median_house_value IS NOT NULL;
```

100 %

Results | Messages

| | PearsonCorrelation |
|---|---|
| 1 | 0.688075207464567 |

```sql
--5. Average Number of Bedrooms Per Household
SELECT
AVG(total_bedrooms/households) AS 'Average number of bedrooms per household'
FROM
housing
WHERE households >0;
```

150 %

Results | Messages

| | Average number of bedrooms per household |
|---|---|
| 1 | 1.09706238580699 |

```sql
--6. Find Areas with High Bedroom-to-Room Ratio
SELECT TOP 10
longitude, latitude, total_bedrooms, total_rooms,
       ROUND(total_bedrooms * 1.0 / total_rooms, 2) AS 'bedroom_room_ratio'
FROM
housing
WHERE total_rooms > 0
ORDER BY bedroom_room_ratio DESC;
```

150 %

Results | Messages

| | longitude | latitude | total_bedrooms | total_rooms | bedroom_room_ratio |
|---|---|---|---|---|---|
| 1 | -121.040000915527 | 37.6699981689453 | 19 | 19 | 1 |
| 2 | -117.790000915527 | 35.2099990844727 | 2 | 2 | 1 |
| 3 | -118.440002441406 | 34.2799987792969 | 11 | 11 | 1 |
| 4 | -118.23999786377 | 34.0400009155273 | 107 | 116 | 0.92 |
| 5 | -121.900001525879 | 37.3699989318848 | 72 | 78 | 0.92 |
| 6 | -118.26000213623 | 34.0499992370605 | 52 | 58 | 0.9 |
| 7 | -118.230003356934 | 34.0499992370605 | 270 | 346 | 0.78 |
| 8 | -114.650001525879 | 32.7900009155273 | 33 | 44 | 0.75 |
| 9 | -121.290000915527 | 37.9500007629395 | 79 | 107 | 0.74 |
| 10 | -121.48999786377 | 38.5800018310547 | 405 | 569 | 0.71 |

```sql
--7. Oldest vs Newest Median Housing Age by Region
SELECT
ocean_proximity,
MAX(housing_median_age) AS 'Oldest',
MIN(housing_median_age) AS 'Newest'
FROM
housing
GROUP BY ocean_proximity;
```

150 %

Results | Messages

| | ocean_proximity | Oldest | Newest |
|---|---|---|---|
| 1 | INLAND | 52 | 1 |
| 2 | NEAR BAY | 52 | 2 |
| 3 | NEAR OCEAN | 52 | 2 |
| 4 | <1H OCEAN | 52 | 2 |
| 5 | ISLAND | 52 | 27 |

```sql
--8. Regions with Median House Value Over $500,000
SELECT TOP 10
ocean_proximity,
median_house_value
FROM
housing
WHERE median_house_value >500000
ORDER BY median_house_value DESC
```

150 %

Results | Messages

| | ocean_proximity | median_house_value |
|---|---|---|
| 1 | <1H OCEAN | 500001 |
| 2 | <1H OCEAN | 500001 |
| 3 | NEAR OCEAN | 500001 |
| 4 | NEAR BAY | 500001 |
| 5 | <1H OCEAN | 500001 |
| 6 | NEAR BAY | 500001 |
| 7 | NEAR OCEAN | 500001 |
| 8 | <1H OCEAN | 500001 |
| 9 | <1H OCEAN | 500001 |
| 10 | <1H OCEAN | 500001 |

```sql
--9. Income Distribution Buckets
SELECT
    CASE
        WHEN median_income < 2 THEN 'Low Income'
        WHEN median_income BETWEEN 2 AND 4 THEN 'Mid Income'
        WHEN median_income BETWEEN 4 AND 6 THEN 'High Income'
        ELSE 'Very High Income'
    END AS income_bracket,
    COUNT(*) AS record_count
FROM housing
GROUP BY
    CASE
        WHEN median_income < 2 THEN 'Low Income'
        WHEN median_income BETWEEN 2 AND 4 THEN 'Mid Income'
        WHEN median_income BETWEEN 4 AND 6 THEN 'High Income'
        ELSE 'Very High Income'
    END;
```

150 %

Results | Messages

| | income_bracket | record_count |
|---|---|---|
| 1 | Low Income | 2439 |
| 2 | Very High Income | 2362 |
| 3 | High Income | 5725 |
| 4 | Mid Income | 10114 |

```sql
--10. Household Size Statistics
SELECT
    MIN(CAST(population AS FLOAT) / NULLIF(households, 0)) AS min_household_size,
    MAX(CAST(population AS FLOAT) / NULLIF(households, 0)) AS max_household_size,
    AVG(CAST(population AS FLOAT) / NULLIF(households, 0)) AS avg_household_size
FROM housing;
```

50 %

Results | Messages

| | min_household_size | max_household_size | avg_household_size |
|---|---|---|---|
| 1 | 0.692307692307692 | 1243.33333333333 | 3.07065515943639 |

## Tools Used

- **Database**: Microsoft SQL Server
- **Query Tool**: SQL Server Management Studio (SSMS)
- **Data Source**: California Housing dataset (CSV)
- **Skills**: SQL (GROUP BY, CASE, aggregate functions, filtering, derived metrics)

## Outcome & Learnings

- Discovered a strong relationship between income levels and house values.
- Identified that proximity to the ocean often correlates with higher house prices.
- Gained insights into urban population density and housing structures.
- Practiced writing complex SQL queries including ratio analysis and manual correlation approximations.