Exploratory Data Analysis

EDA Insight

From the file.csv.gz extraction, i found another few file that is:

- **California Housing**
  Our first glimpse takes us to the California housing dataset, a treasure trove of information about housing in the Golden State. Through detailed statistical summaries, we aim to uncover patterns in housing prices, explore relationships between variables, and understand the factors influencing the real estate market.

- **Mnist**
  Next on our exploration is the MNIST dataset, a classic in the realm of machine learning. This collection of handwritten digits challenges us to delve into the world of image recognition. We'll seek to understand the distribution of digits, identify patterns, and perhaps uncover nuances in the art of digit penmanship.

- **Anscombe**
  An unexpected yet fascinating addition to our dataset is Anscombe's quartet—a set of four datasets that share statistical properties but differ significantly when graphically displayed. This enigma challenges us to question the reliance on summary statistics alone and underscores the importance of visualization in data exploration

EDA Insight (Cont.)

Because of the large variety of different files, I decided to use the **California Housing** file because:

- ***Real-World Relevance***: The California housing dataset encapsulates real-world dynamics of one of the most dynamic real estate markets.
- ***Societal Impact***: Housing is not just about numbers; it's about homes, communities, and the heartbeat of society.
- ***Practical Applications***: Insights derived from this dataset can have practical applications in real estate planning, urban development, and policy making.
- ***Richness of Variables***: The California housing dataset encompasses a variety of variables—from location-specific features to housing characteristics.

After this there will be some analysis of the **dataset** I chose

Data Overview

Before we embark on the detailed exploration of the **California Housing** data, let's first gain a comprehensive understanding of our dataset. The data overview provides a snapshot, offering insights into the nature and characteristics of the information we're about to delve into.

Key Highlights:

- Variable Types and Uniqueness
- Missing Values
- Statistical Summaries
- Central Tendencies and Spread

Data Overview (Cont.)

	Data Types	Unique Values	Missing Values	Mean	Median	Min	25th Percentile	50th Percentile (Median)	75th Percentile	Max
longitude	float64	607	0	-119.589200	-118.48500	-124.1800	-121.810	-118.48500	-118.020000	-114.4900
latitude	float64	587	0	35.635390	34.27000	32.5600	33.930	34.27000	37.690000	41.9200
housing_median_age	float64	52	0	28.845333	29.00000	1.0000	18.000	29.00000	37.000000	52.0000
total_rooms	float64	2215	0	2599.578667	2106.00000	6.0000	1401.000	2106.00000	3129.000000	30450.0000
total_bedrooms	float64	1055	0	529.950667	437.00000	2.0000	291.000	437.00000	636.000000	5419.0000
population	float64	1802	0	1402.798667	1155.00000	5.0000	780.000	1155.00000	1742.750000	11935.0000
households	float64	1026	0	489.912000	409.50000	2.0000	273.000	409.50000	597.250000	4930.0000
median_income	float64	2578	0	3.807272	3.48715	0.4999	2.544	3.48715	4.656475	15.0001
median_house_value	float64	1784	0	205846.275000	177650.00000	22500.0000	121200.000	177650.00000	263975.000000	500001.0000

Data Overview (Cont.)

- **Variable Types and Uniqueness:**

Our California housing dataset is characterized by various data types, predominantly featuring floating-point values. Among the key variables, we observe that latitude and longitude are continuous numerical variables, while median income, housing median age, and other features also fall within this category. The datasets richness is evident in the substantial number of unique values, signifying the diversity captured in variables such as latitude, longitude, and median income.

- **Missing Values:**

A meticulous examination reveals that our dataset is remarkably clean, with no missing values across all variables. This completeness ensures that our analyses are grounded in a comprehensive dataset, allowing us to draw meaningful insights without the need for imputation or correction.

- **Statistical Summaries**:

Diving into statistical summaries, we uncover the central tendencies and distributions of our variables. From the mean and median values, we gain insights into the typical or central values of features like median income, housing median age, and more. The minimum and maximum values provide bounds, revealing the range within which most data points lie.

- **Central Tendencies**:

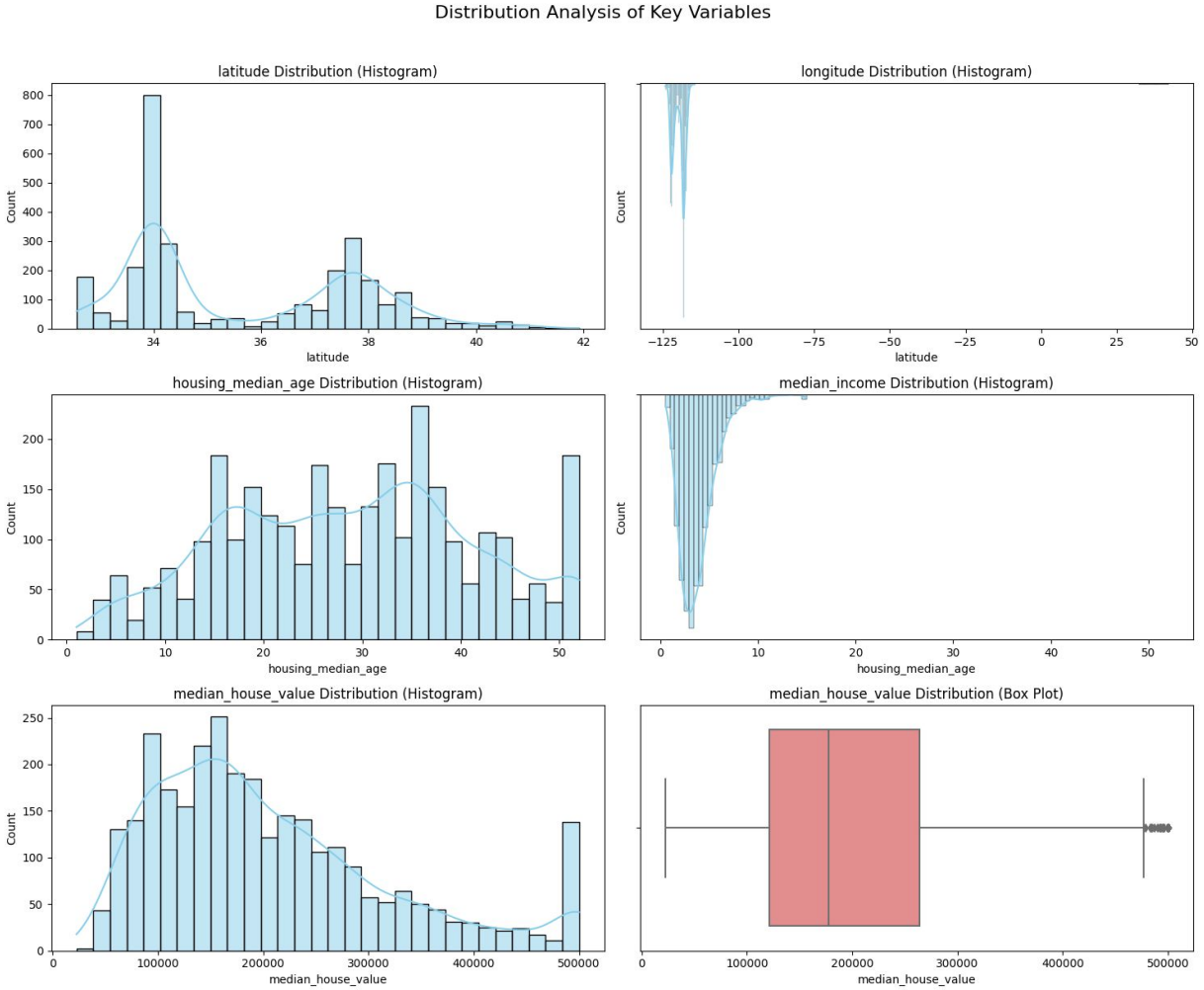
Examining central tendencies further emphasizes the typical values within our dataset. The mean and median values give us a sense of the average or central point for each variable. For instance, the mean median income and median median income shed light on the overall income distribution in the California housing dataset.

- **Spread Measures:**

Understanding the spread of our data is crucial for grasping the variability and distribution. Minimum and maximum values illustrate the range covered by our variables, while the standard deviation provides a measure of how much individual data points deviate from the mean. This knowledge sets the stage for identifying outliers and understanding the dispersion within our dataset.

In conclusion, the California housing dataset stands as a robust foundation for exploration, offering diverse variables, completeness, and a rich landscape of insights waiting to be uncovered.

Distribution Analysis



Distribution Analysis

As we transition from the **overview** to a more granular analysis, our focus now turns to the **distribution of key variables** within the California housing dataset. Understanding these distributions is crucial for unraveling **patterns**, identifying **outliers**, and **gaining insights** into the underlying structure of our data.

Latitude and Longitude:
Latitude Distribution: The distribution of latitude values spans from a minimum of 32.56 to a maximum of 41.92, suggesting a diverse representation of locations across California. Visualizing this distribution could reveal geographic clusters or patterns related to housing dynamics.

Longitude Distribution: The distribution of longitude values, ranging from -124.18 to -114.49, echoes the expansive geographical coverage of our dataset. Analyzing this distribution may unveil regional variations and their impact on housing features.

Distribution Analysis (Cont.)

Housing Median Age:
The distribution of housing median age, with a minimum of 1 and a maximum of 52, showcases the diversity in the age of housing units. A visualization of this distribution might highlight areas with newer or older housing stock, influencing market dynamics.

Income Metrics:
Median Income Distribution: The distribution of median income, with a mean of 3.807 and a median of 3.487, offers insights into the income landscape. A skewed distribution might indicate areas with varying economic profiles, impacting housing values and affordability.

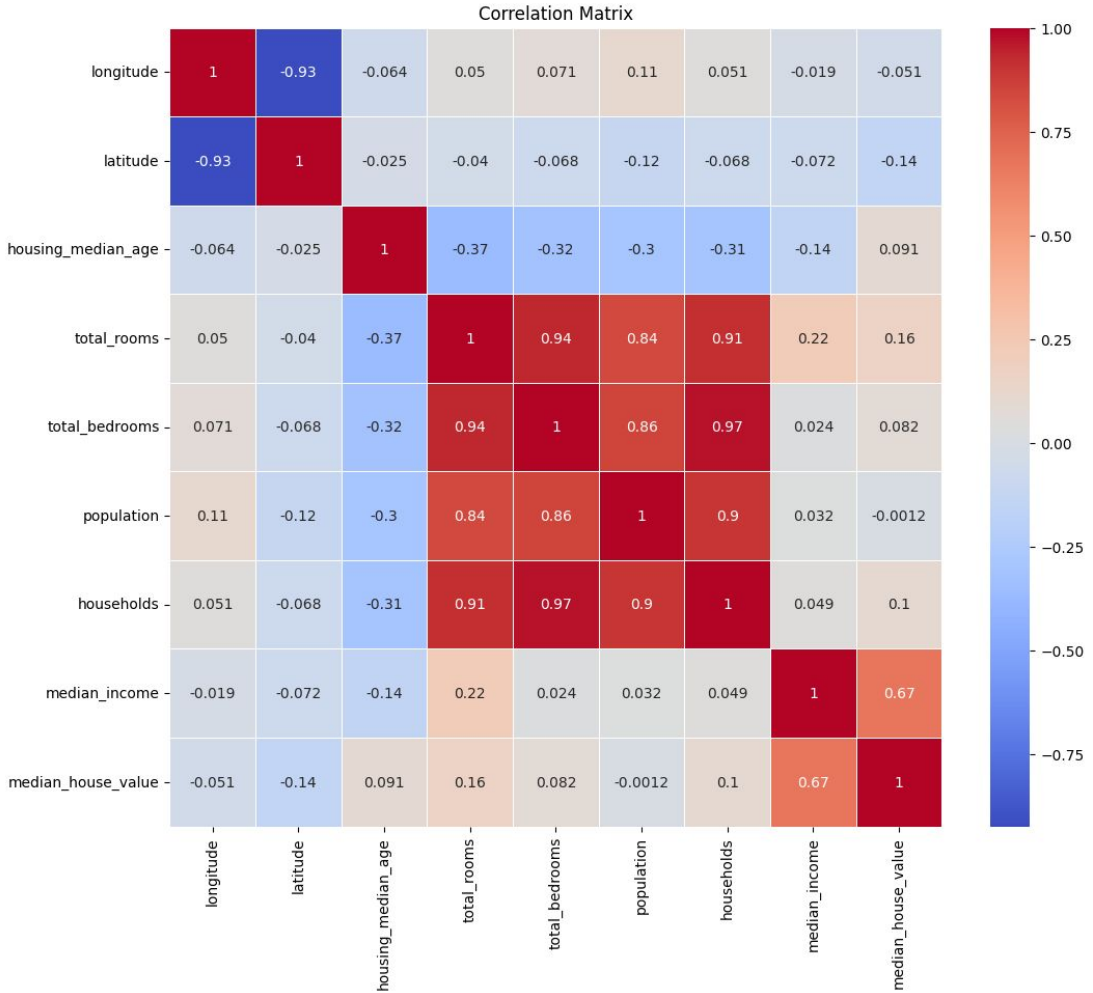
Housing Characteristics:
Total Rooms, Bedrooms, Population, and Households: Exploring the distributions of these variables—total rooms, total bedrooms, population, and households—provides a deeper understanding of the housing landscape. Skewed or symmetrical distributions can unveil characteristics of different neighborhoods, influencing housing demand and supply.

Distribution Analysis (Cont.)

Median House Value:
The distribution of median house values, ranging from $22,500 to $500,001, serves as a key indicator of the housing market. Analyzing this distribution can reveal areas with higher or lower property values, contributing to the broader narrative of real estate dynamics.

Conclusion:
As we immerse ourselves in the distribution analysis, each histogram and density plot becomes a window into the nuances of California's housing story. From the spatial dynamics captured by latitude and longitude to the economic fabric depicted by income distributions, our exploration deepens, promising a tapestry of insights waiting to be uncovered.

Correlation Analysis



Correlation Analysis

Moving beyond individual variable distributions, our exploration now shifts to understanding the relationships between key variables within the California housing dataset. Correlation analysis is a powerful tool that allows us to unveil connections, dependencies, and potential influencing factors.

Latitude and Longitude:
Spatial Relationships: Exploring the correlation between latitude and longitude unveils spatial relationships within our dataset. A close examination might reveal patterns such as coastal influences or mountainous terrain, influencing housing dynamics in specific regions.

Housing Characteristics:
Total Rooms, Bedrooms, Population, and Households: Analyzing the interplay between these variables offers insights into the internal dynamics of housing units. Strong correlations might indicate dependencies, such as a higher number of rooms correlating with a higher population.

Economic Factors:
Median Income and Median House Value: Understanding the correlation between income levels and housing values is crucial for gauging affordability. A positive correlation could signify that areas with higher incomes tend to have higher property values.

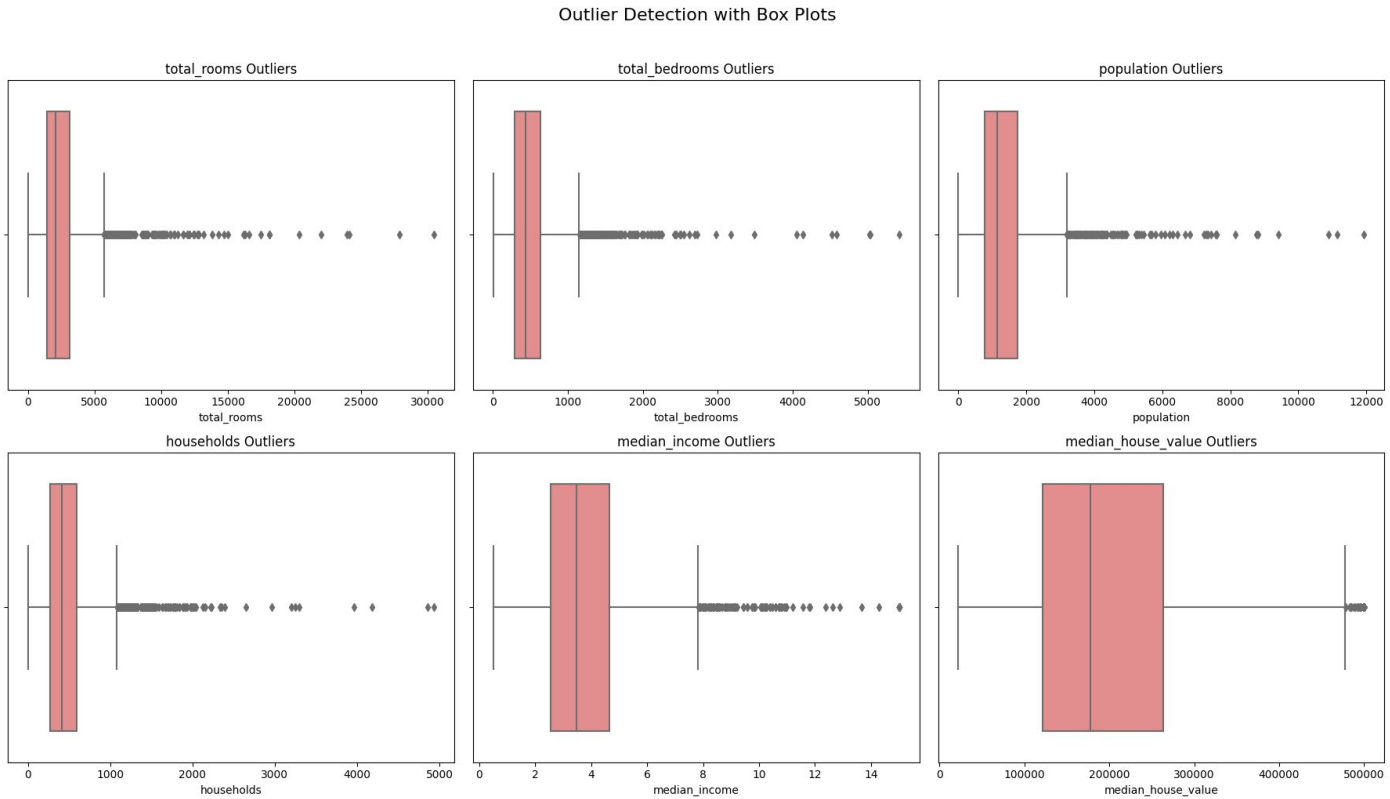
Correlation Analysis

Housing Median Age:
Influence on Housing Characteristics: Exploring how housing median age correlates with other features, such as total rooms or bedrooms, can highlight the impact of historical development patterns on housing structures.

Insights from the Correlation Matrix:
Our correlation matrix encapsulates these relationships numerically, ranging from -1 to 1. A value close to 1 implies a strong positive correlation, while a value close to -1 indicates a strong negative correlation. Values near 0 suggest weak or no correlation.

Conclusion:
As we delve into the intricate web of correlations, each coefficient becomes a thread in the narrative of California's housing story. The relationships we uncover will guide us in understanding the multifaceted influences that shape the real estate landscape across the state.

Outlier Detection



Outliers Detection

As we navigate the intricate landscape of the California housing dataset, our attention turns to identifying potential outliers—data points that deviate significantly from the majority. The detection of outliers is crucial for refining our analyses, ensuring the integrity of our insights, and gaining a nuanced understanding of the data distribution.

Outliers in Housing Characteristics:
Total Rooms, Bedrooms, Population, and Households: Employing box plots for these variables, we can visually pinpoint extreme values that may represent unusual housing structures or population concentrations. Outliers in these features could signal unique properties or neighborhoods that deviate from the norm.

Economic Factors and Housing Values:
Median Income and Median House Value: Analyzing box plots for income levels and housing values helps us identify potential outliers that might influence our understanding of affordability and property valuation. Extreme values in these variables could signify areas with distinct economic characteristics.

Outliers Detection

Spatial Outliers:
Latitude and Longitude: Spatial outliers might manifest as extreme latitude and longitude values. Investigating these points can unveil geographic anomalies that may impact housing dynamics, such as remote locations or areas with exceptional geographical features.t.

Conclusion:
As we embark on the quest to detect outliers, each data point beyond the norm becomes a beacon guiding us to hidden nuances within the dataset. These outliers unveil stories of uniqueness, influencing the broader narrative of California's housing landscape.

Key Finding

- The California housing dataset is diverse, encompassing variables related to spatial information, housing characteristics, economic factors, and housing values.
- Exploration revealed spatial patterns, potential correlations, and areas with unique housing dynamics.
- Outlier detection highlighted specific data points that deviate significantly from the norm, potentially indicating areas of interest or anomalies.