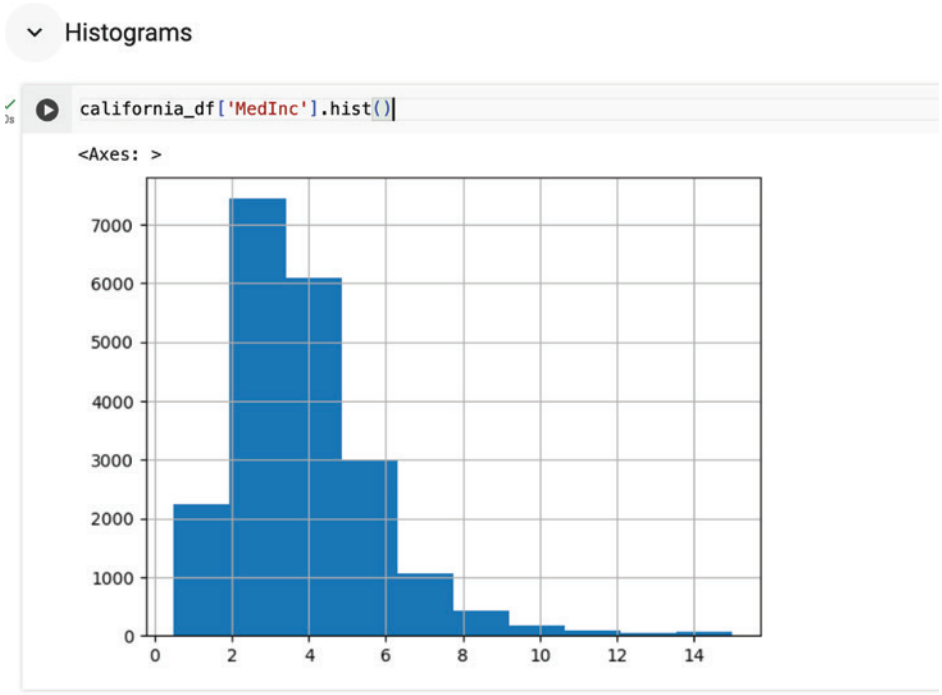
Chapter 12

Univariate Analysis: Histograms and Skew

Histograms are graphical representations of the frequency distribution of numerical variables. They display the frequency of data values within predefined intervals or bins, allowing analysts to visualize the shape and spread of the data distribution. Histograms are key for: visualization of distribution, identifying patterns, summary of data range, detection of outliers, comparison between groups, and data transformation. Overall, histograms are a powerful tool in exploratory data analysis because they provide a quick and intuitive way to understand the characteristics and distribution of a single variable in your dataset.

12.1 Basic Histograms with Pandas

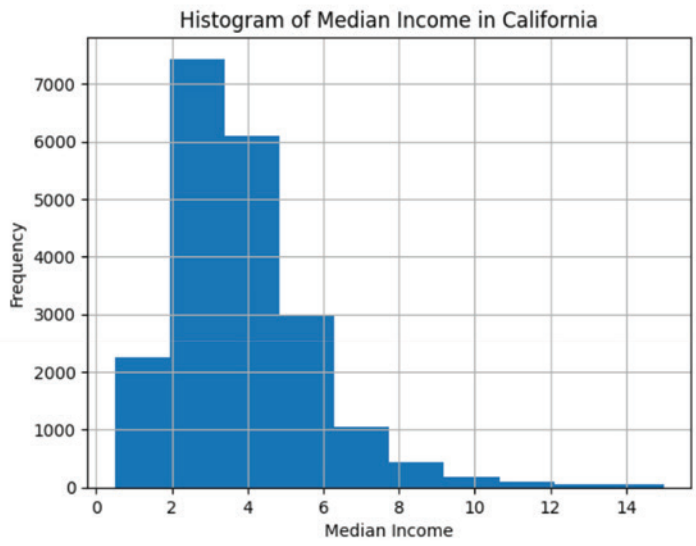


Using the hist() attribute, a histogram can be created using pandas. Whenever charts are built, it is key to note that a title, x-label, and y-label should be included. Let's add code, so now the plot is more clear. We will now be using the Matplotlib library to plot histograms with further customizations.

12.2 Customized Histograms with Matplotlib

Clearer histograms with Matplotlib

```
import pandas as pd
import matplotlib.pyplot as plt

california_df['MedInc'].hist()

plt.title('Histogram of Median Income in California')
plt.xlabel('Median Income')
plt.ylabel('Frequency')
plt.show()
```



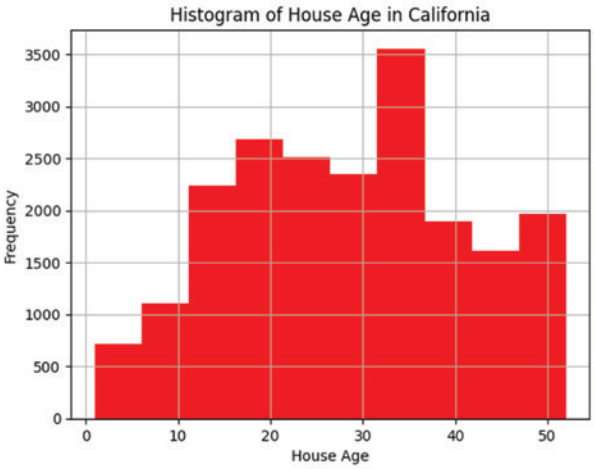
This plot is a histogram with MedInc. The distribution of MedInc is relatively normal (from Chapter 6), and there seem to be a few outliers after 10.

By changing the column name for each variable, we can create different histograms. Each histogram can also be customized by color, size, x-label ticks, and so on.

12.3 Plotting Remaining Histograms with Matplotlib

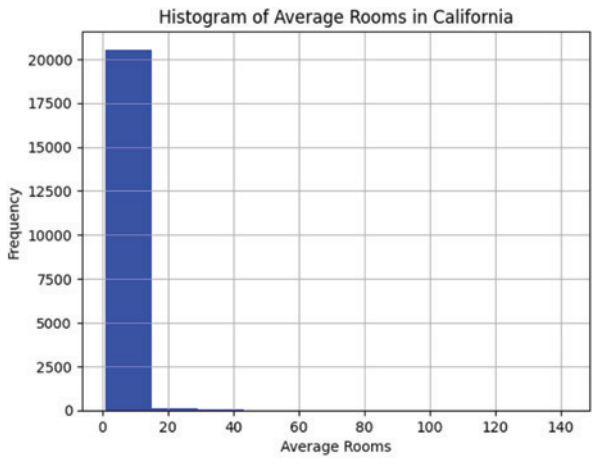
HouseAge

```
import pandas as pd
import matplotlib.pyplot as plt

california_df['HouseAge'].hist(color='red')

plt.title('Histogram of House Age in California')
plt.xlabel('House Age')
plt.ylabel('Frequency')

plt.show()
```



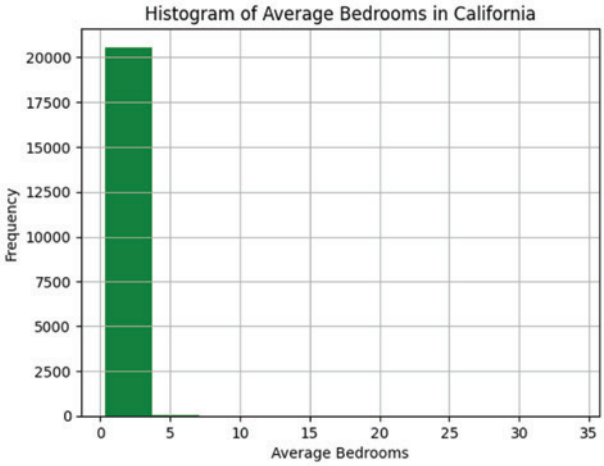
AveRooms

```
import pandas as pd
import matplotlib.pyplot as plt

california_df['AveRooms'].hist(color='blue')

plt.title('Histogram of Average Rooms in California')
plt.xlabel('Average Rooms')
plt.ylabel('Frequency')

plt.show()
```



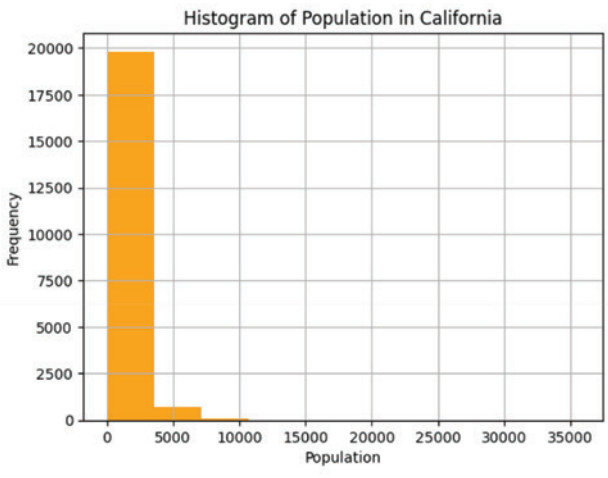
AveBedrms

```
import pandas as pd
import matplotlib.pyplot as plt

california_df['AveBedrms'].hist(color='green')

plt.title('Histogram of Average Bedrooms in California')
plt.xlabel('Average Bedrooms')
plt.ylabel('Frequency')

plt.show()
```



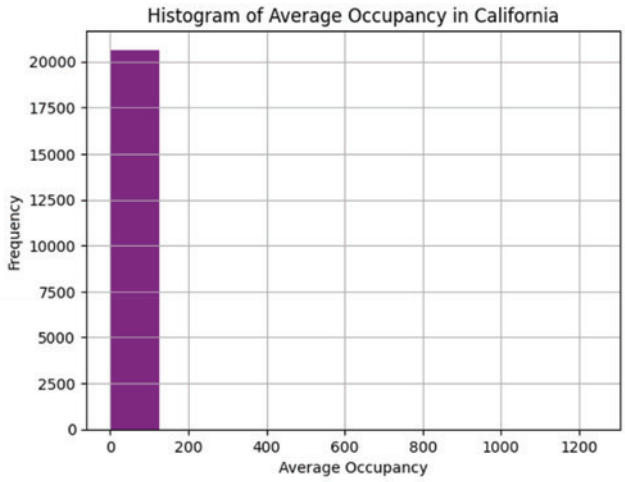
Population

```
import pandas as pd
import matplotlib.pyplot as plt

california_df['Population'].hist(color='orange')

plt.title('Histogram of Population in California')
plt.xlabel('Population')
plt.ylabel('Frequency')

plt.show()
```



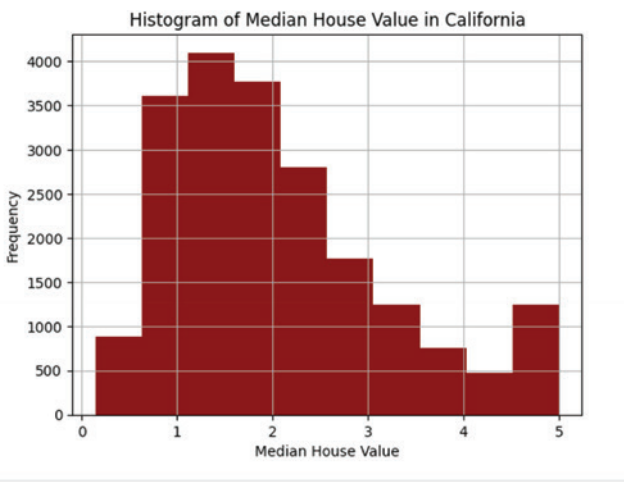
AveOccup

```
import pandas as pd
import matplotlib.pyplot as plt

california_df['AveOccup'].hist(color='purple')

plt.title('Histogram of Average Occupancy in California')
plt.xlabel('Average Occupancy')
plt.ylabel('Frequency')

plt.show()
```



MedHouseVal

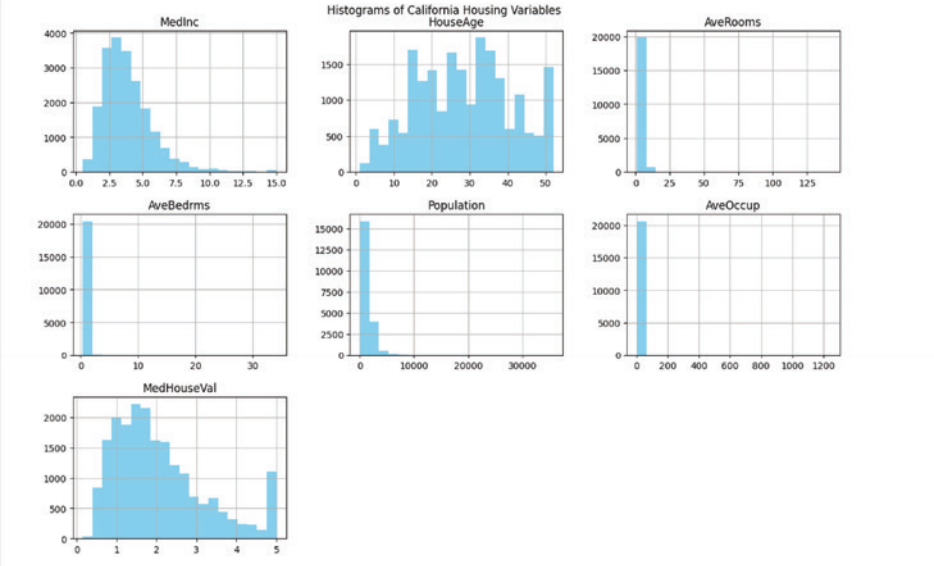
```
import pandas as pd
import matplotlib.pyplot as plt

california_df['MedHouseVal'].hist(color='darkred')

plt.title('Histogram of Median House Value in California')
plt.xlabel('Median House Value')
plt.ylabel('Frequency')

plt.show()
```



12.4 All Histograms

A few easy lines of code to generate all the histograms are also available.

```
california_df.hist(figsize=(15, 10), color='skyblue', bins=20)
plt.suptitle('Histograms of California Housing Variables', x=0.5, y=0.92)
plt.show()
```



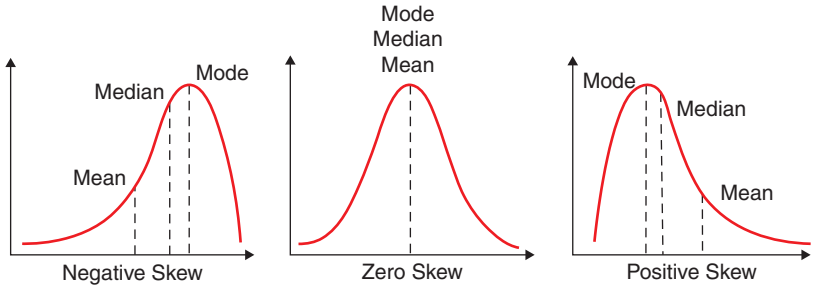
12.5 Interpretation

- MedInc: The variable has a normal distribution. The standard deviation is small as the observations are close to each other in a tighter range. There are a few noticeable outliers after the value 10. These are points in the dataset with incomes that are very high and could be explained by the outliers.
- HouseAge: This variable has a normal distribution as well. However, the standard deviation is very large, as indicated by the distribution being wider and flatter. There are multiple modes at points 25, 15, and 35.
- AveRooms: This variable has an exponential decay distribution. Most of the points are close to 0, and there are few outliers after the 0 values. This plot is difficult to read as the outliers are skewing the distribution of smaller numbers.
- AveBedrms: This variable has an exponential decay distribution. Most of the points are close to 0, and there are few outliers after the 0 values. This plot is difficult to read as the outliers are skewing the distribution of smaller numbers.
- Population: This variable has an exponential decay distribution. Most of the points are close to 0, and there are few outliers after the 0 values to 10,000. This plot is difficult to read as the outliers are skewing the distribution of smaller numbers.
- AveOccup: This variable has an exponential decay distribution. Most of the points are close to 0, and there are few outliers after the 0 values. This plot is difficult to read as the outliers are skewing the distribution of smaller numbers.
- MedHouseVal: This variable has a normal distribution as well. However, the standard deviation is very large, as indicated by the distribution being wider and flatter. There is also a noticeable skew.

There are data points that are getting closer to the end on the right. This is a right-skew distribution, which makes sense as there are fewer houses that have a higher Median House Value.

12.6 Skew

Symmetric distribution: A symmetric distribution is one where the data is evenly distributed around the mean, with roughly equal tails on both sides of the distribution. For example, a normal distribution is symmetric.

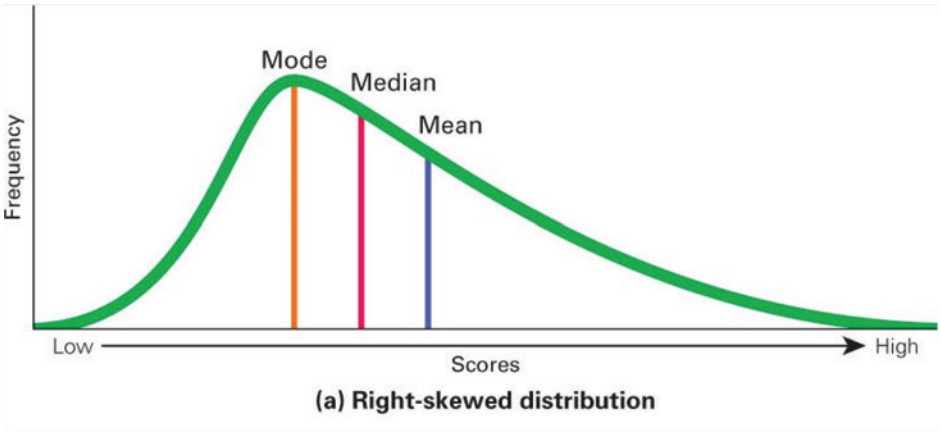


Skewness: Skewness measures the degree of asymmetry in a distribution.

12.7 Positive Skew/Right Skew

Positive Skewness (Right Skew): If the right tail (larger values) of the distribution is longer or fatter than the left tail, the distribution is said to be positively skewed. This means that there are more data points on the left side of the distribution with fewer extreme values on the right side. In a histogram, this appears as a longer tail to the right.

In a positively skewed distribution, the mean will be greater than the median, and the mode will be the smallest value.

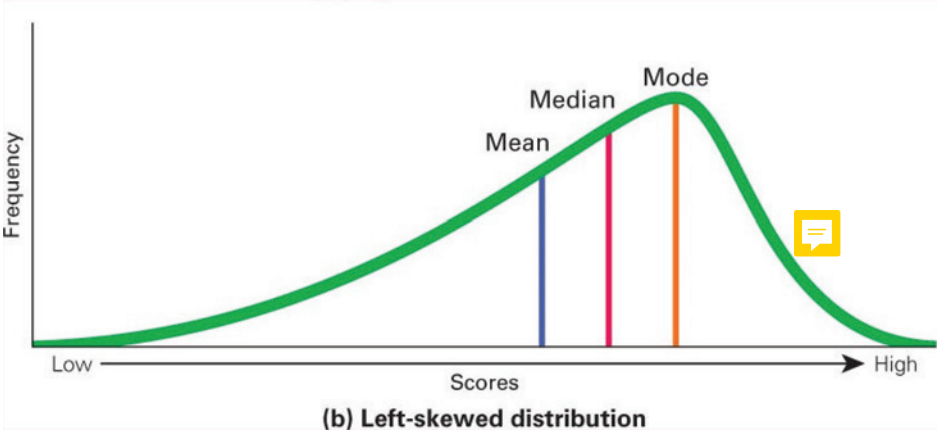


12.8 Negative Skew/Left Skew

Negative Skewness (Left Skew): If the left tail (smaller values) of the distribution is longer or fatter than the right tail, the distribution is said to be negatively skewed. This means that there are more data points on the

right side of the distribution with fewer extreme values on the left side. In a histogram, this appears as a longer tail to the left.

In a negatively skewed distribution, the mean will be less than the median, and the mode will be the largest value.



(b) Left-skewed distribution

12.9 Conclusion

In conclusion, along with descriptive statistics, it is important to understand graphically the impact that skewness has on the variables. Descriptive statistics provide valuable numerical summaries of the data, including measures of central tendency and dispersion. However, graphical representations such as histograms or density plots offer visual insights into the distribution of the data. Skewness, which measures the asymmetry of the distribution, can have a significant impact on our understanding of the variables. By visually examining the skewness of the data, we can better grasp how the data is distributed, whether it is heavily concentrated on one side, and how it deviates from a symmetric pattern. This graphical understanding complements the numerical measures, providing a more comprehensive view of the dataset and aiding in the interpretation of the results.