ECON 0150 | Economic Data Analysis

The economist's data analysis pipeline.

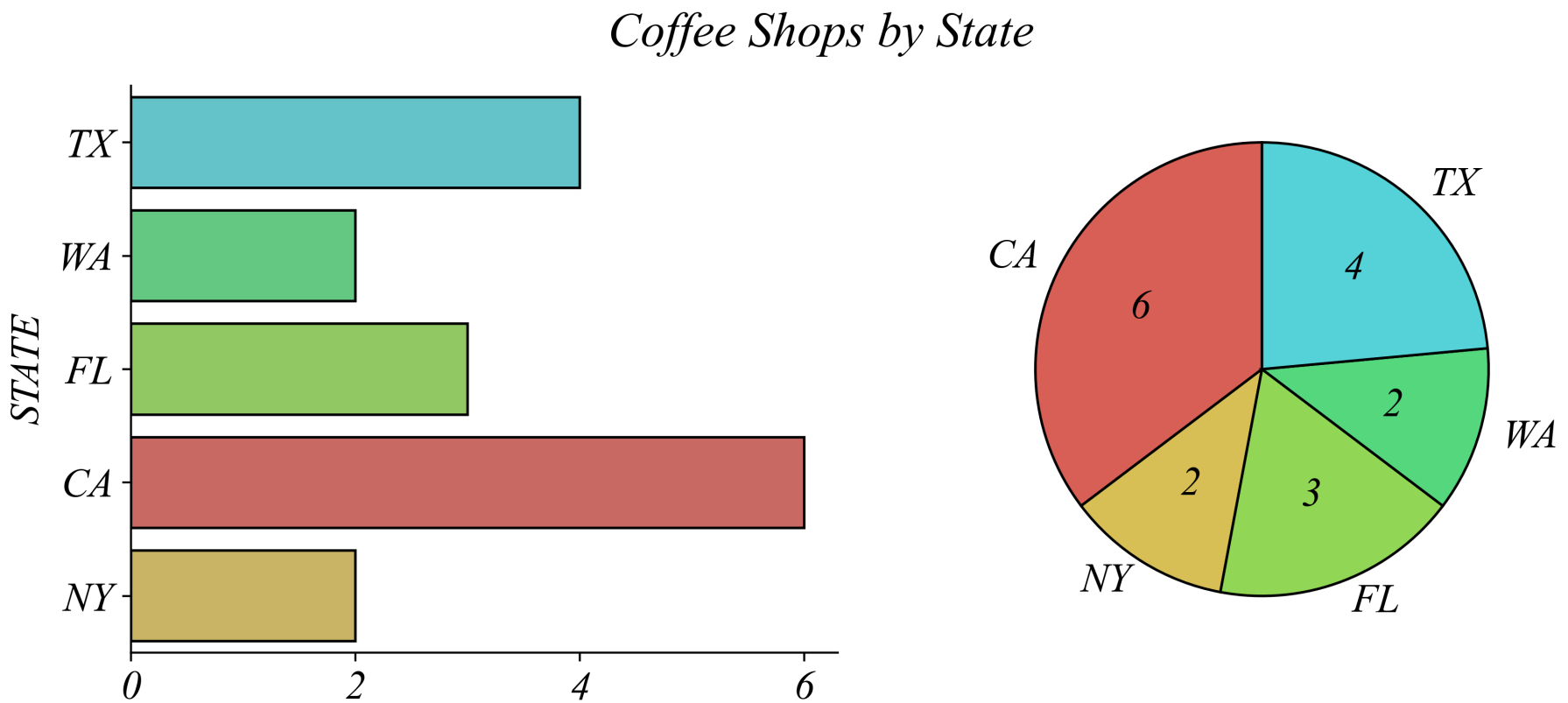
Part 1.1 | Summarizing Categorical Variables

Summarizing Categorical Variables

... use the appropriate summary tool for the variable type

Catagorical Variables: Visualizations

Q. Which state has the most locations?

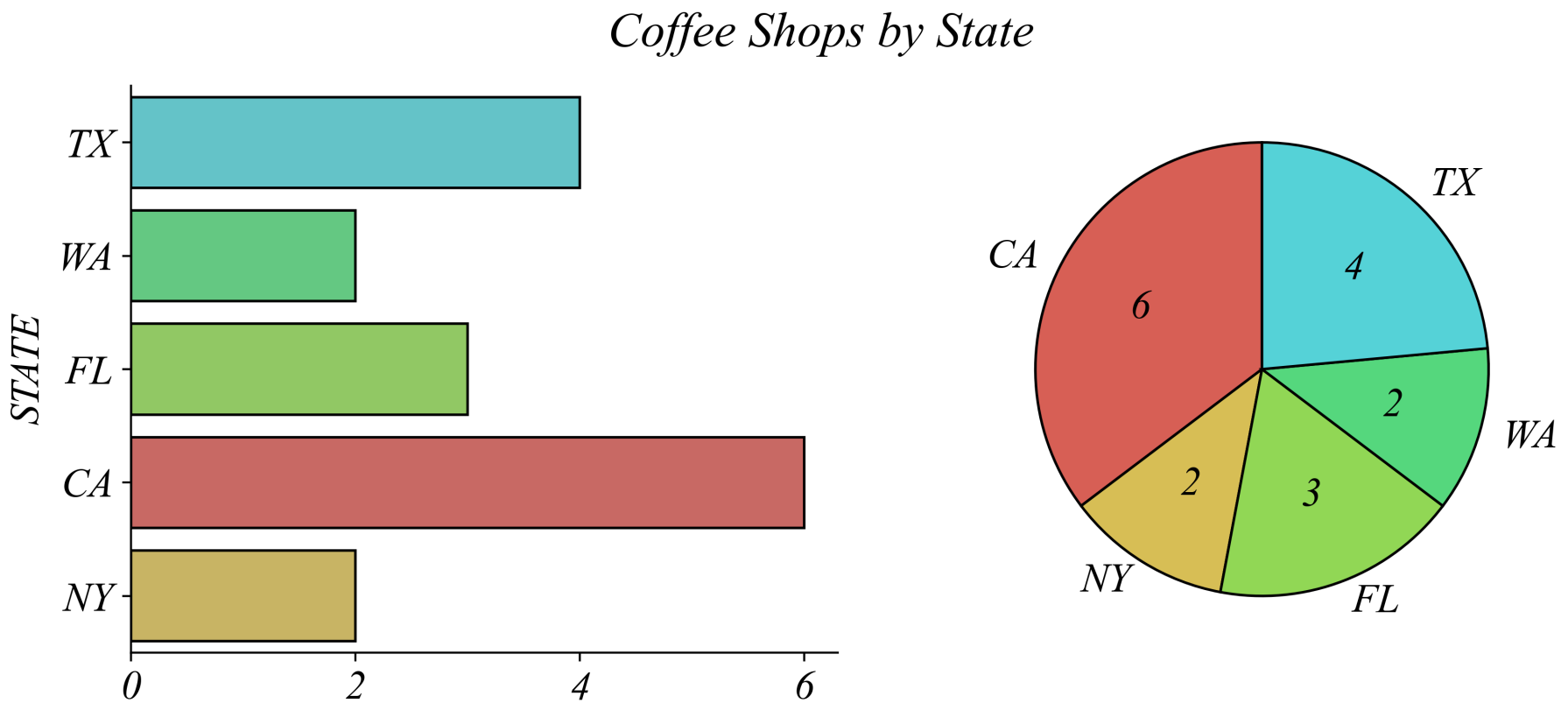


> *pay attention to which of these two figures is easier to answer the question*

> *it's pretty easy to see that it's CA from both of these figures*

Catagorical Variables: Visualizations

Q. Does FL or WA have more shops?

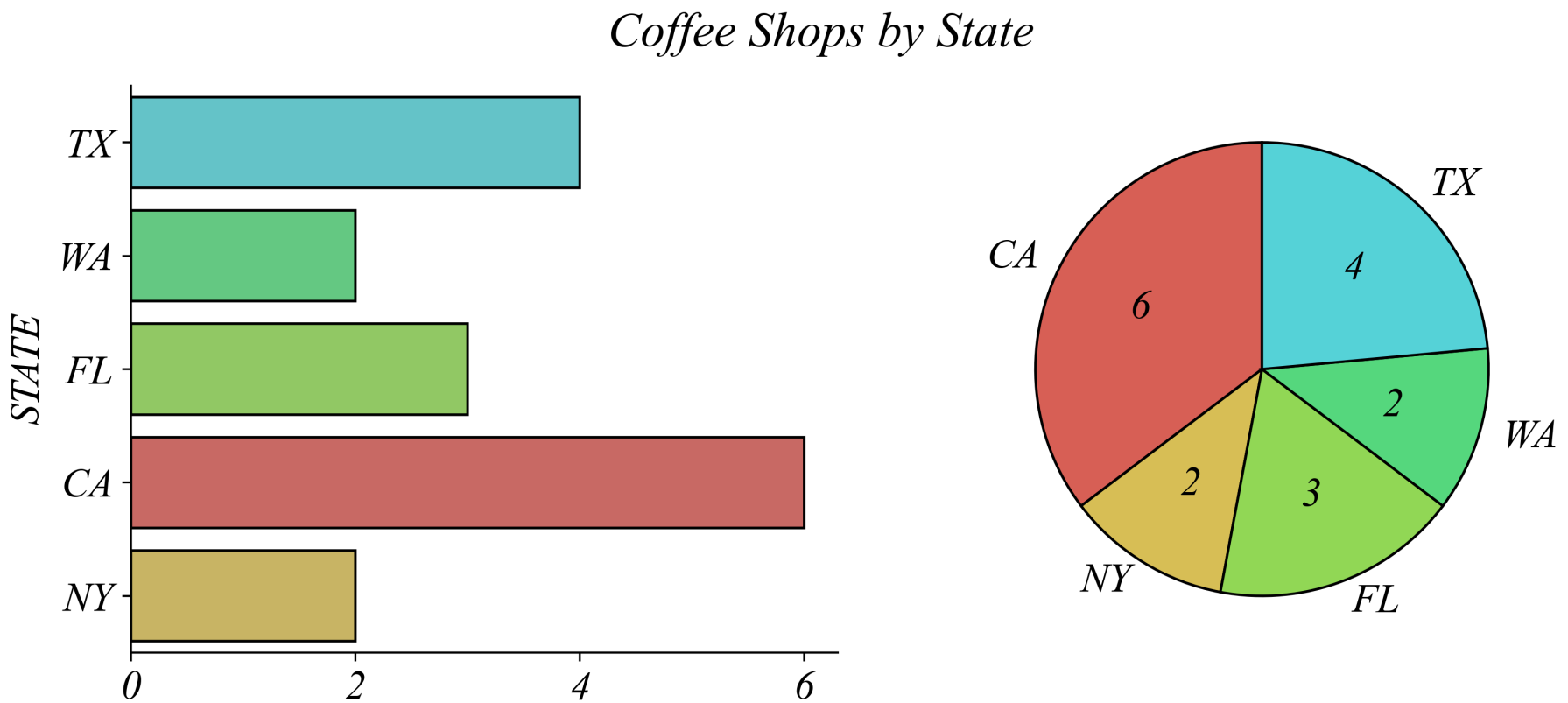


> *pay attention to which of these two figures is easier to answer the question*

> *a bar graph is much easier to read*

Catagorical Variables: Visualizations

Q. How many shops are in FL?

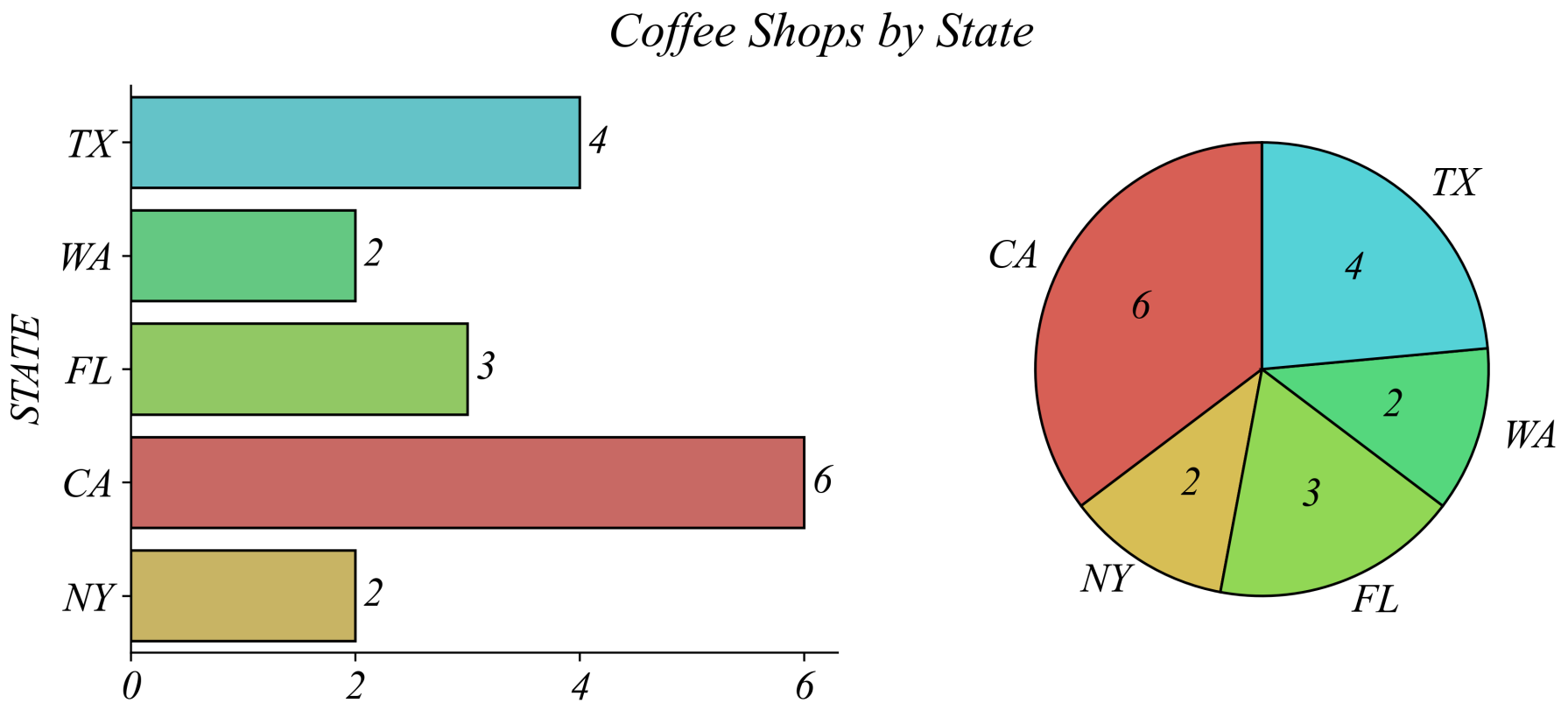


> *pay attention to which of these two figures is easier to answer the question*

> *now it takes a second to read the bar graph…*

Catagorical Variables: Bar Plots

Q. How many shops are in FL?

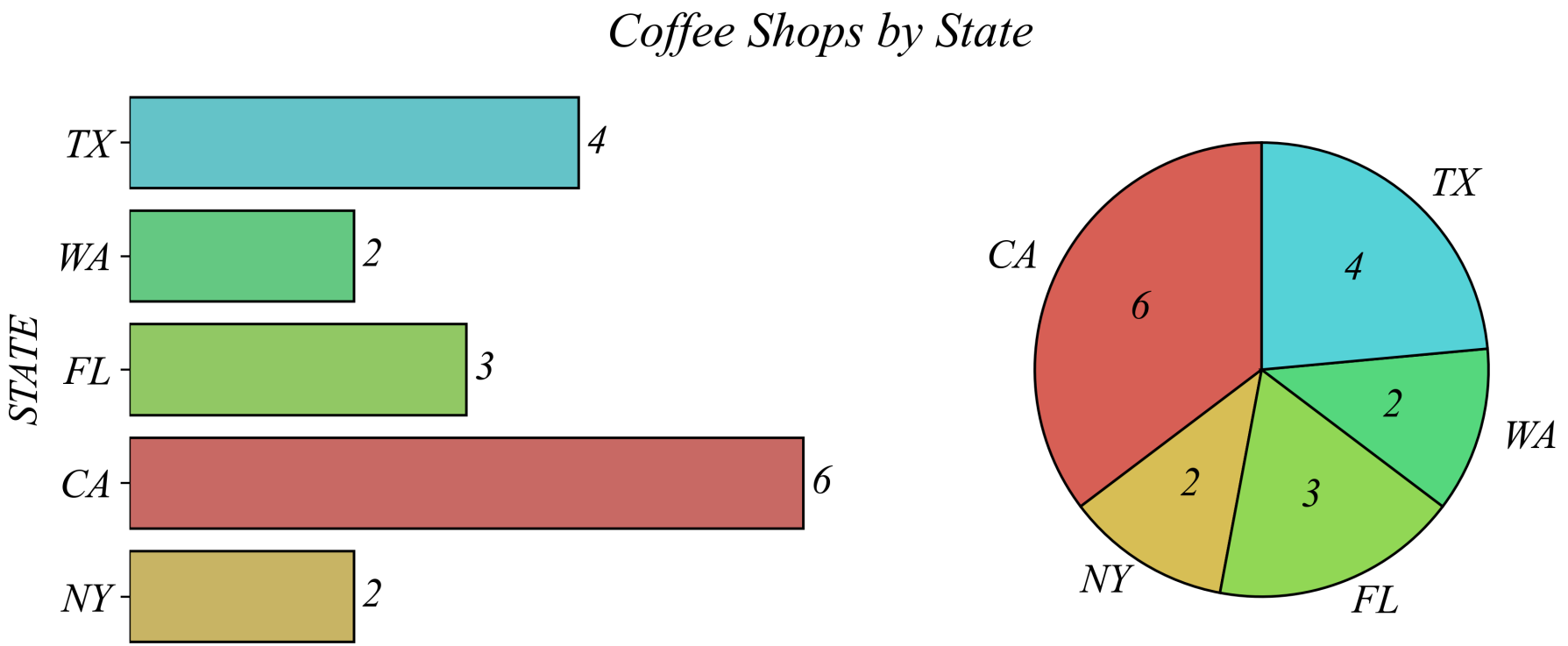


> *pay attention to which of these two figures is easier to answer the question*

> *we can make the bar graph easier to read by placing the number near the bar*

Catagorical Variables: Remove Clutter

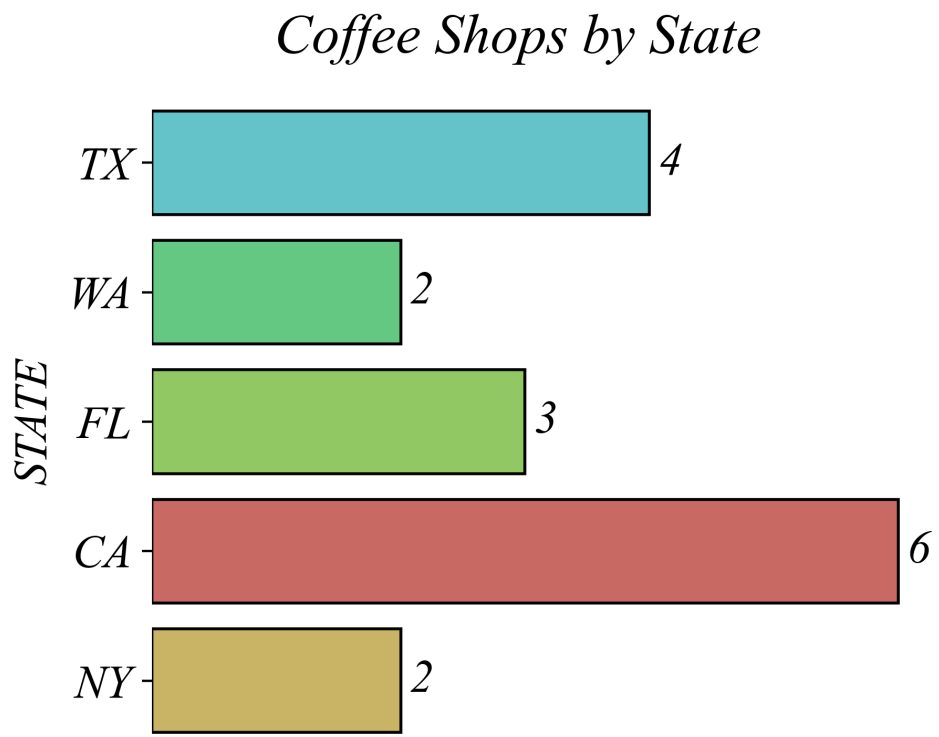
Q. How many shops are in the state with the second most locations?



> removing clutter guides your eye to the important information

Catagorical Variables: Remove Clutter

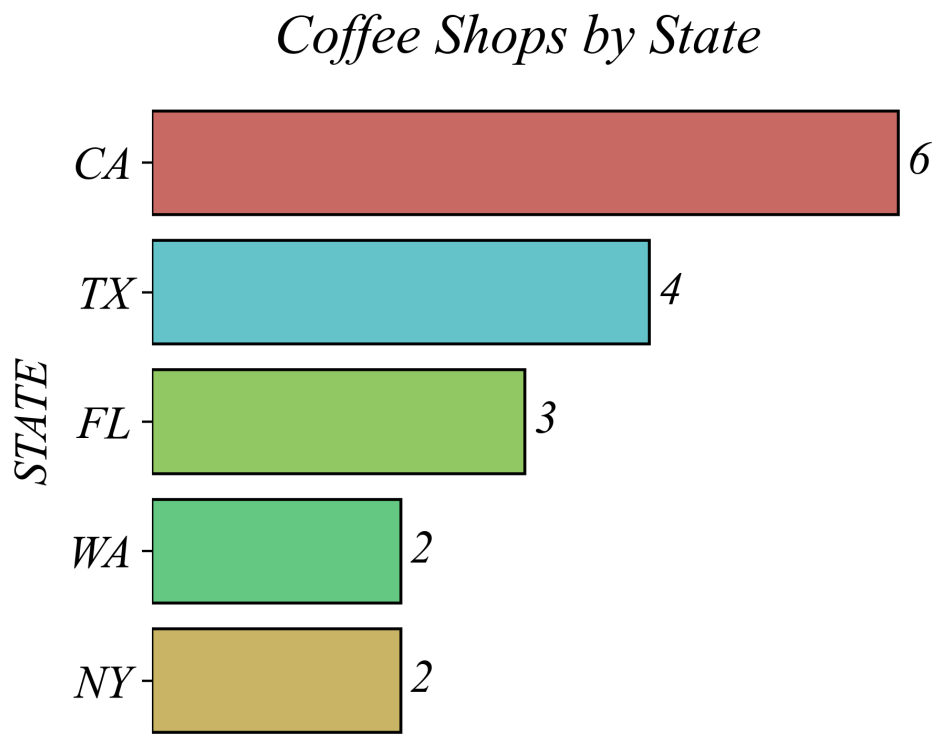
Q. How many shops are in the state with the second most locations?



> removing clutter guides your eye to the important information

Catagorical Variables: Order by Size

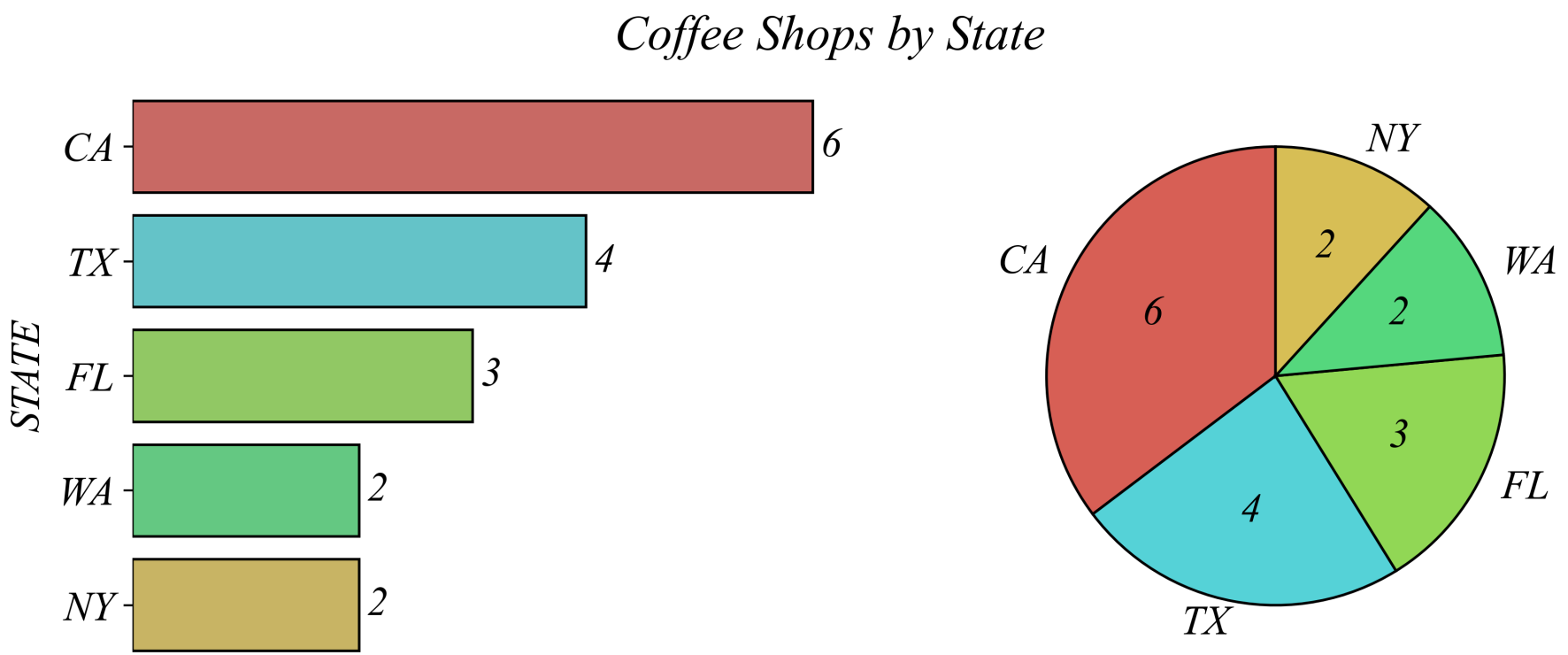
Q. How many shops are in the state with the second most locations?



> states have no inherent order, but sorting can make comparisons easier

Binary Categorical Variables: CA vs Other

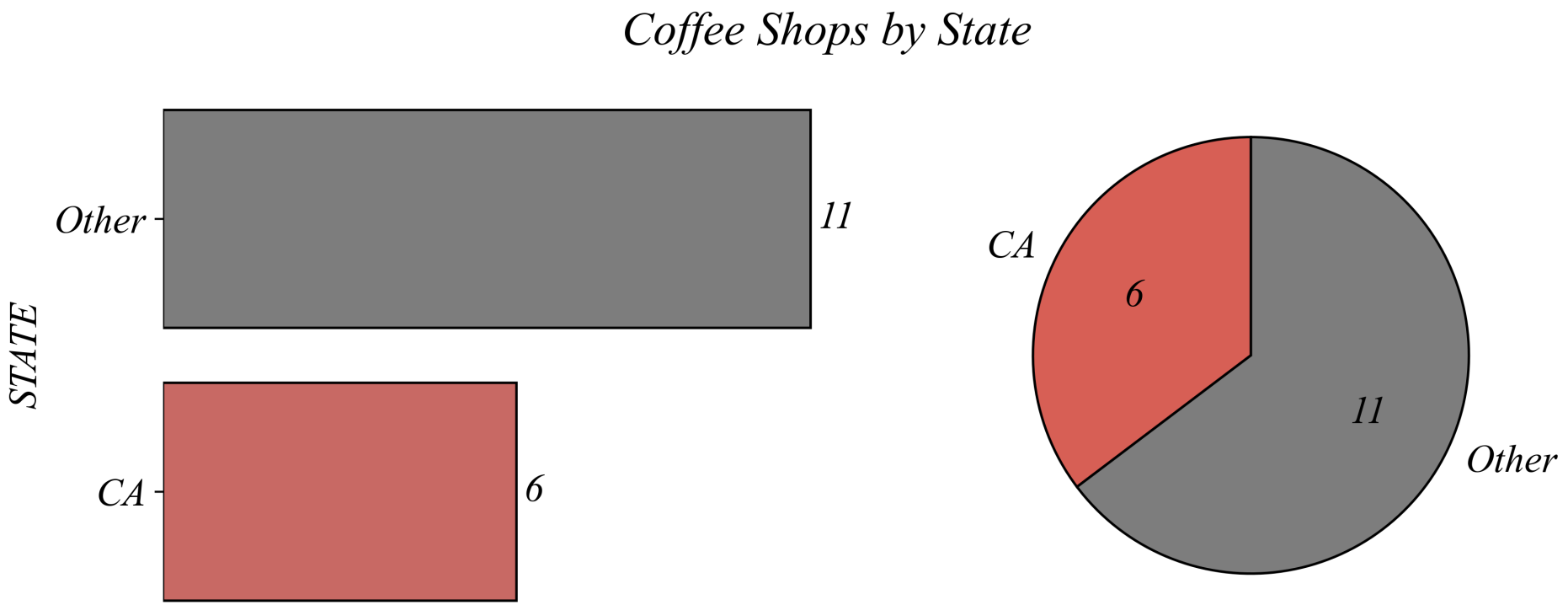
Q. How does CA compare to the whole?



> instead of a nominal categorical variable, this is binary (CA / Other)

Binary Categorical Variables: Binary Visualization

Q. How does CA compare to the whole?

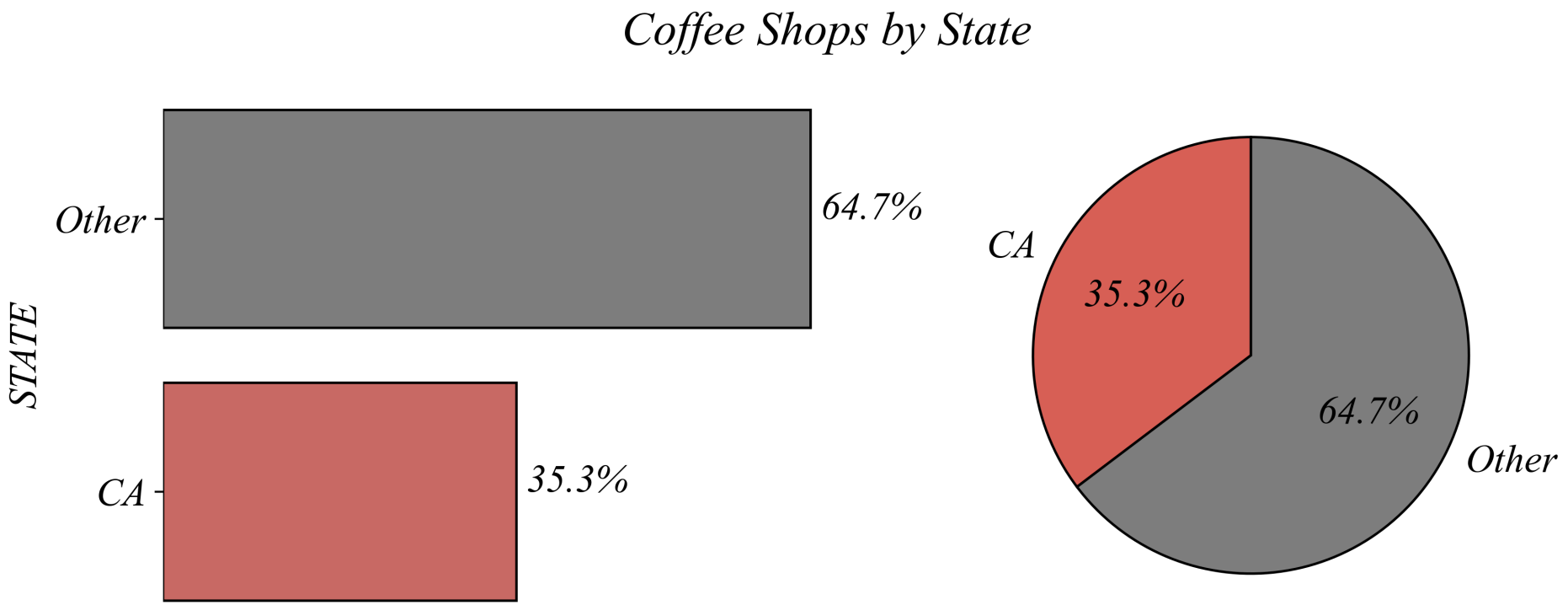


> this question is much easier to see when visualizing the two categories

> here both the pie and the bar communicte the data effectively

Binary Categorical Variables: Percentages

Q. How does CA compare to the whole?



> if the question is about percentages, a pie chart may work best

Takeaways

... use the right summary tool for the variable type

- *Binary Categorical Variables: use a **pie chart** or **bar graph***
- *Nominal Categorical Variables: use a **bar graph**; maybe order by value*
- *Ordinal Categorical Variables: use an **ordered bar graph***
- *Remove clutter; keep it simple*
- *Place information near the object it describes*

Exercises 1.1 | Categorical Variables

Lets visualize coffee shops by state.

- ***Dataset 1:*** `Coffee_Shops.csv`

Lets visualize the main variable in each dataset.

- ***Dataset 2:*** `employment_status.csv`
- ***Dataset 3:*** `household_savings.csv`
- ***Dataset 4:*** `household_incomes.csv`

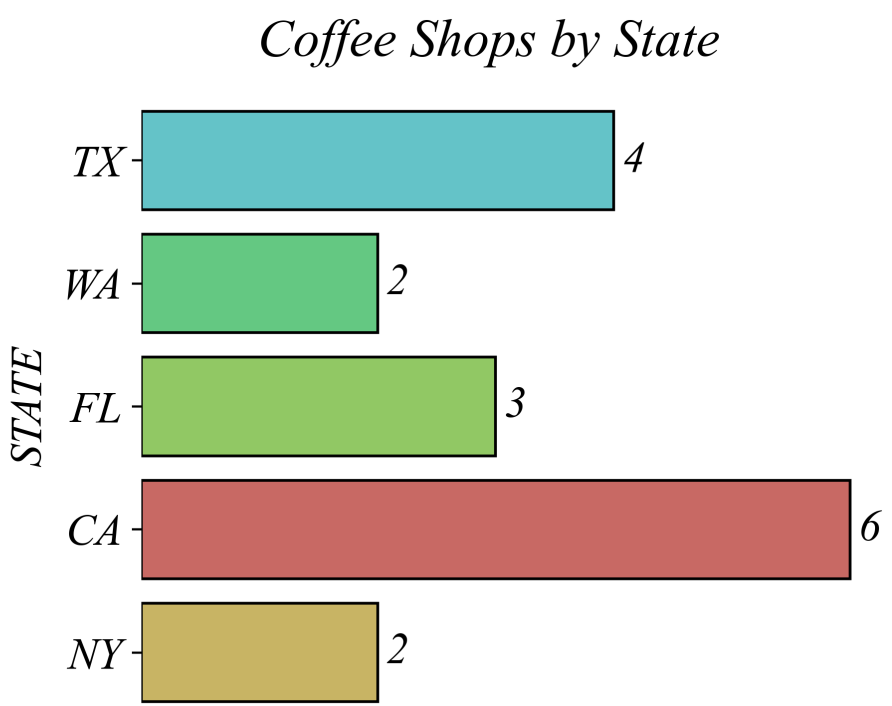
Exercise: Dataset 1

Summarize `Coffee_Shops.csv` as a nominal categorical variable.

```
# Load Dataset
shops = pd.read_csv(file_path + 'Coffee_Shops.csv')
```

```
# Summary Table
shops.value_counts()
```

```
# Countplot (bar plot)
sns.countplot(data=shops, y='STATE', hue='STATE')
```



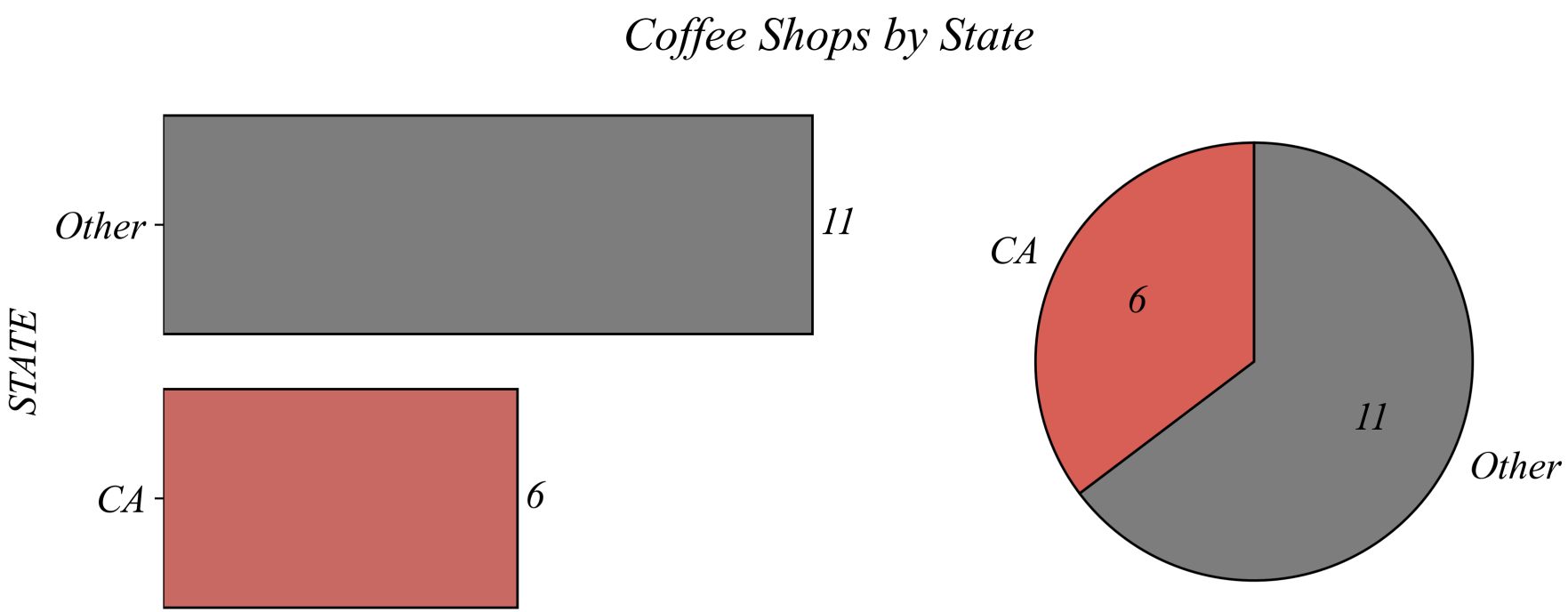
Exercise: Dataset 1

Summarize `Coffee_Shops.csv` as a binary categorical variable.

```
# Load Dataset
shops = pd.read_csv(file_path + 'Coffee_Shops.csv')
```

```
# Create a binary categorical variable
shops['CA'] = np.where(shops['STATE'] == 'CA', 'CA', 'Other')
```

```
# Countplot
sns.countplot(data=shops, y='CA', hue='CA')
```



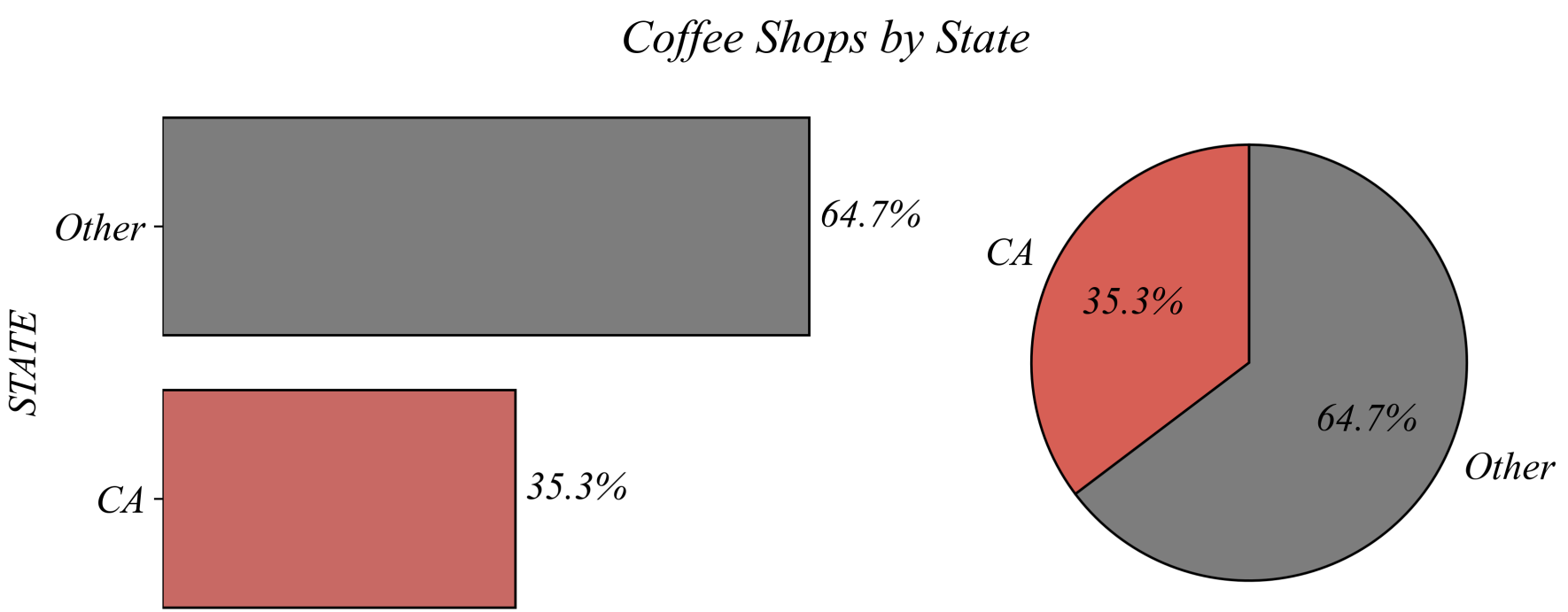
Exercise: Dataset 1

Summarize `Coffee_Shops.csv` as a binary categorical variable.

```
# Load Dataset
shops = pd.read_csv(file_path + 'Coffee_Shops.csv')
```

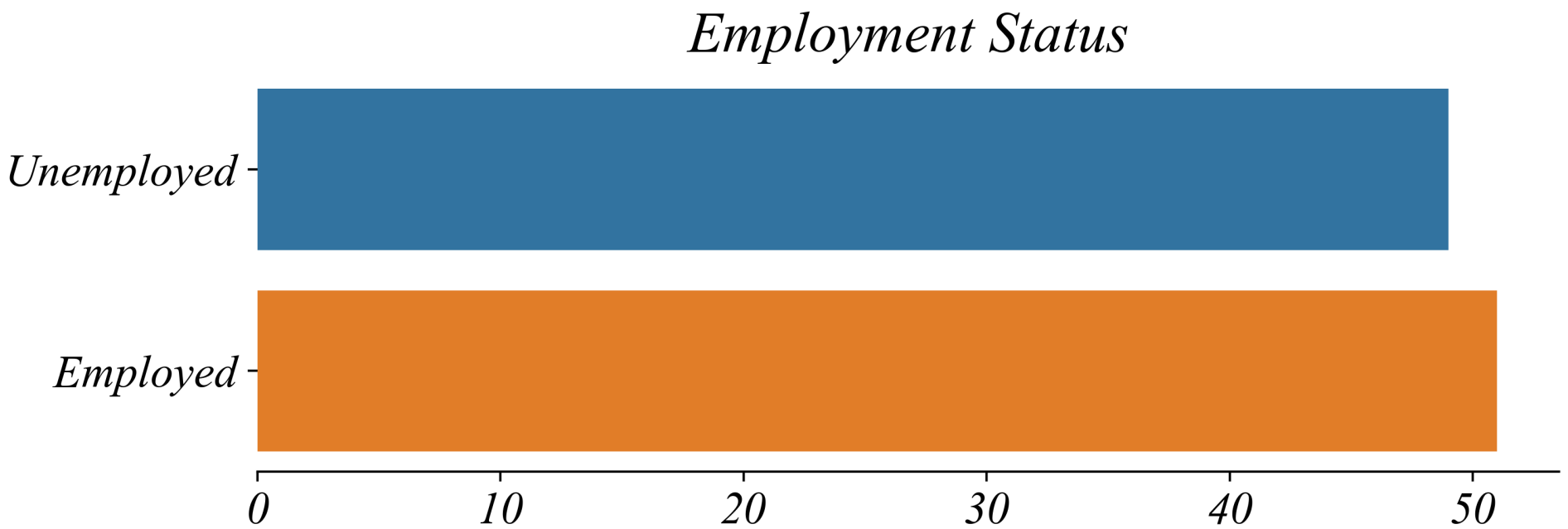
```
# Create a binary categorical variable
shops['CA'] = np.where(shops['STATE'] == 'CA', 'CA', 'Other')
```

```
# Pie Chart
shops['CA'].value_counts().plot(kind='pie', autopct='%1.1f%%')
```



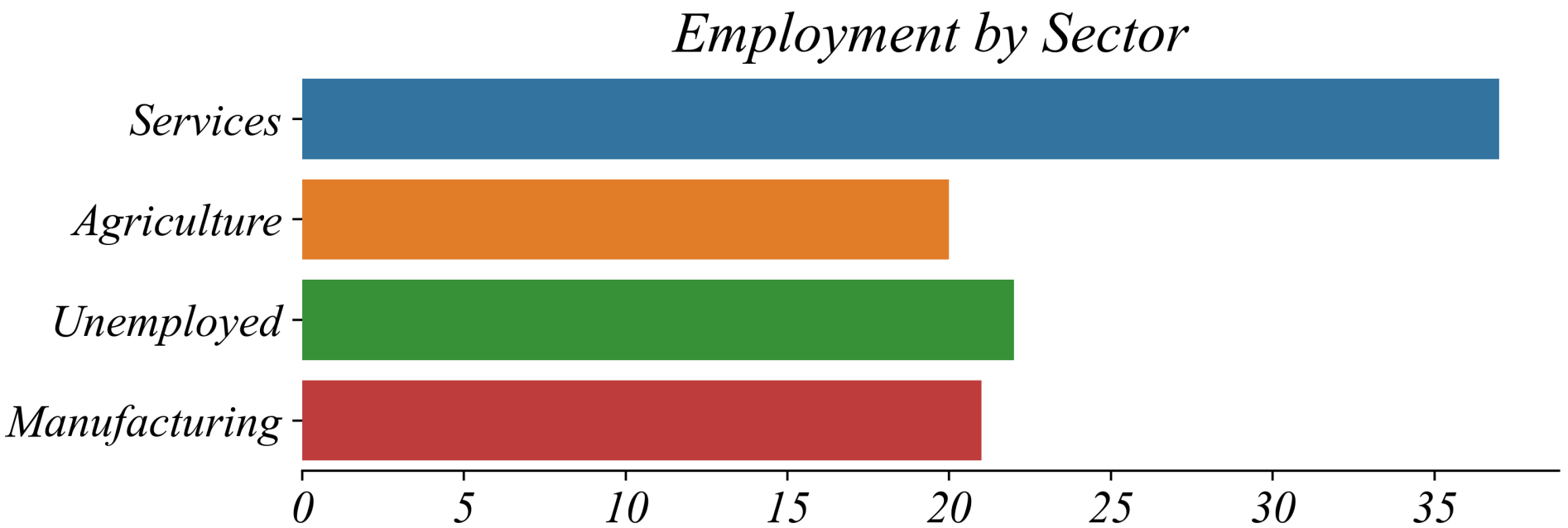
Exercise: Dataset 2

Summarize `employment_status.csv`.



Exercise: Dataset 3

Summarize `household_savings.csv`.



Exercise: Dataset 4

Summarize `household_incomes.csv`.