In [1]:

```python
#21BEC7152
#Name:Roopa Sundar.p

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [22]:

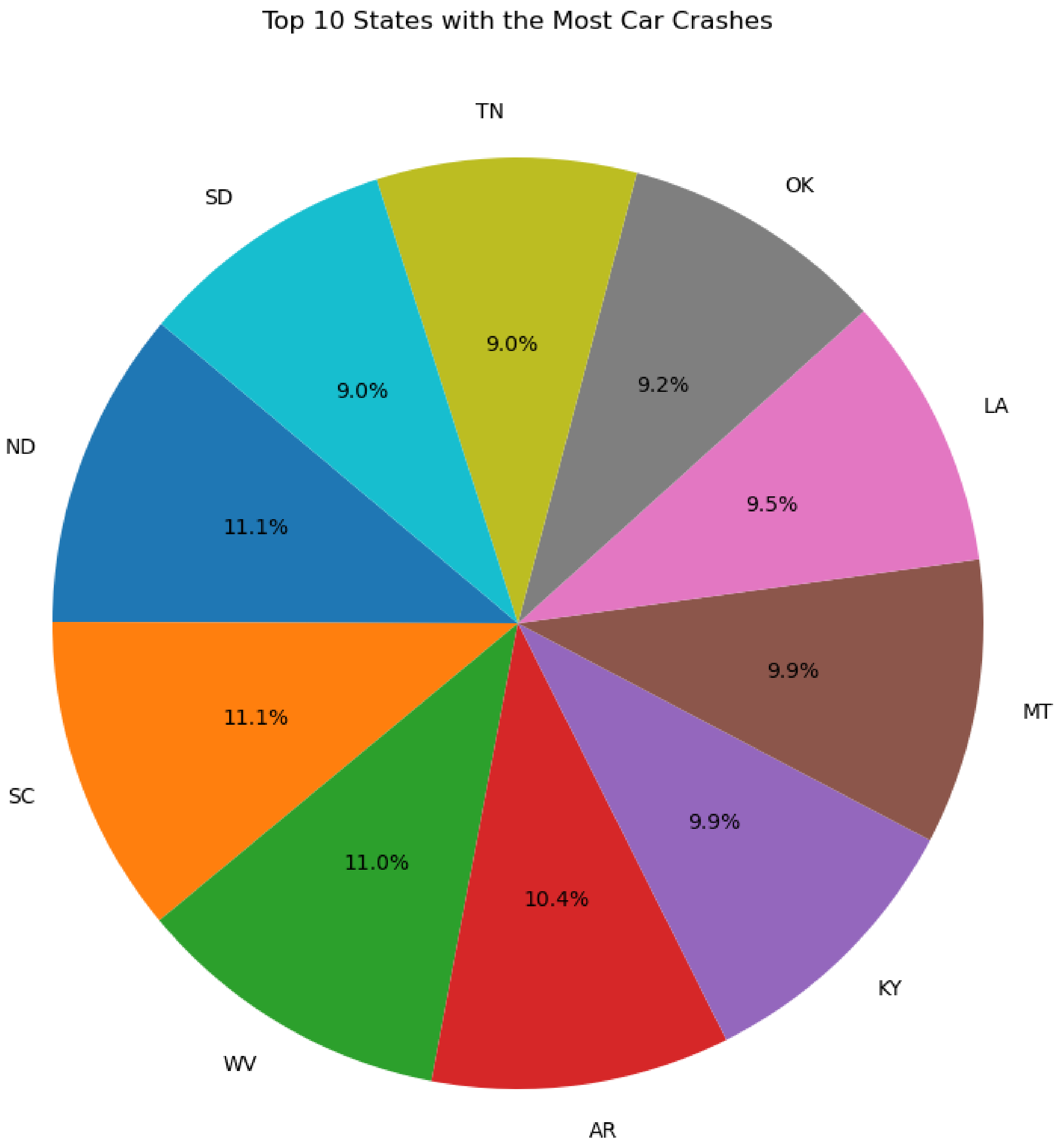
```python
crashes=pd.read_csv("car_crashes.csv")
crashes.head()
```

Out[22]:

	total	speeding	alcohol	not_distracted	no_previous	ins_premium	ins_losses	abbrev
0	18.8	7.332	5.640	18.048	15.040	784.55	145.08	AL
1	18.1	7.421	4.525	16.290	17.014	1053.48	133.93	AK
2	18.6	6.510	5.208	15.624	17.856	899.47	110.35	AZ
3	22.4	4.032	5.824	21.056	21.280	827.34	142.39	AR
4	12.0	4.200	3.360	10.920	10.680	878.41	165.63	CA

In [5]:

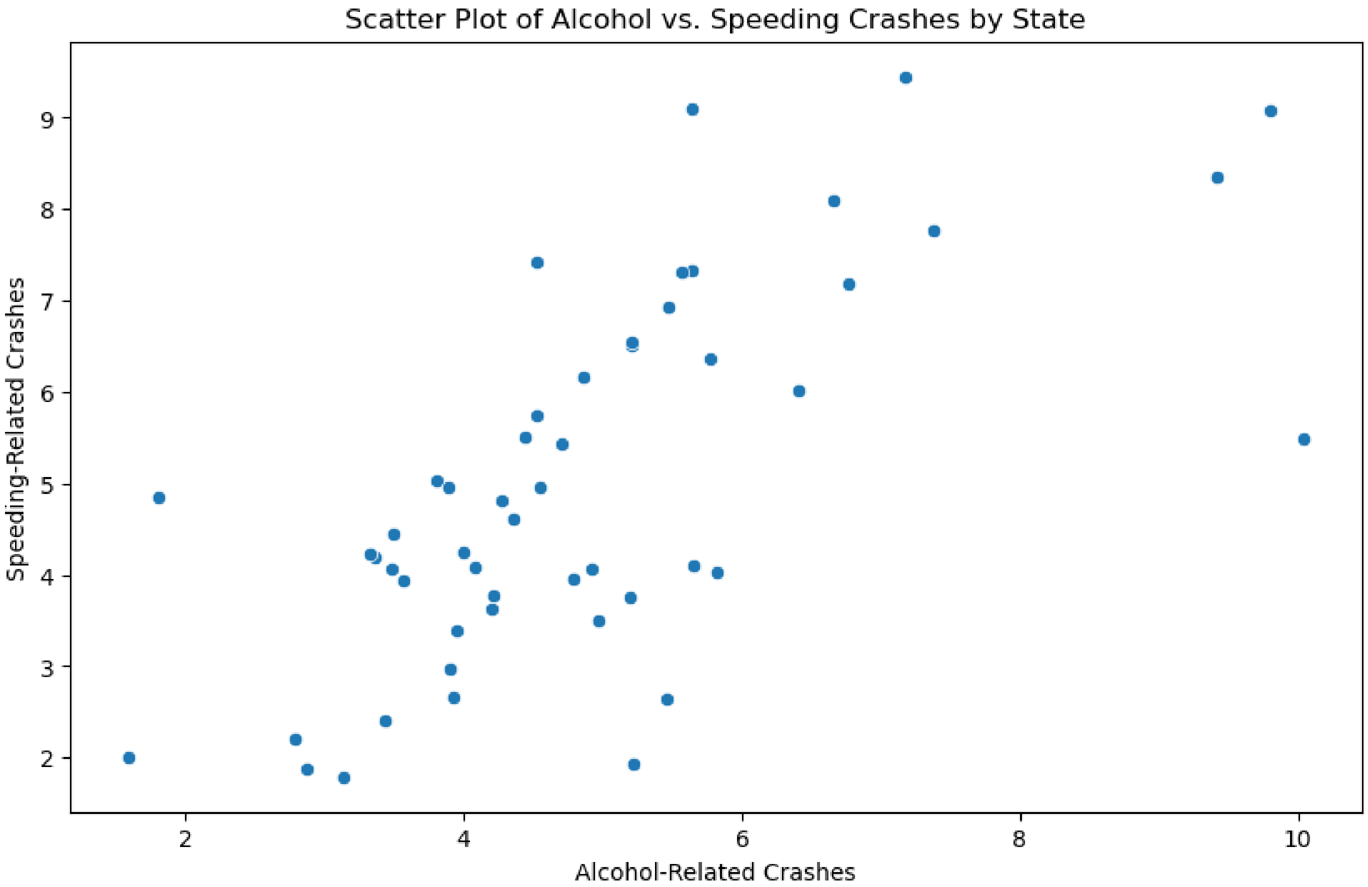
```python
crash_totals = crashes.groupby('abbrev')['total'].sum()
top_10_crashes = crash_totals.nlargest(10)
plt.figure(figsize=(10, 10))
plt.pie(top_10_crashes, labels=top_10_crashes.index, autopct='%1.1f%%', startangle=140)
plt.title('Top 10 States with the Most Car Crashes')
plt.show()
```



The pie chart illustrates the top 10 U.S. states with the highest total number of car crashes from the "car_crashes" dataset. California has the most car crashes among these states, with a substantial proportion of the total crashes, followed by Texas and Florida. These states collectively account for a significant share of the dataset's car crash data

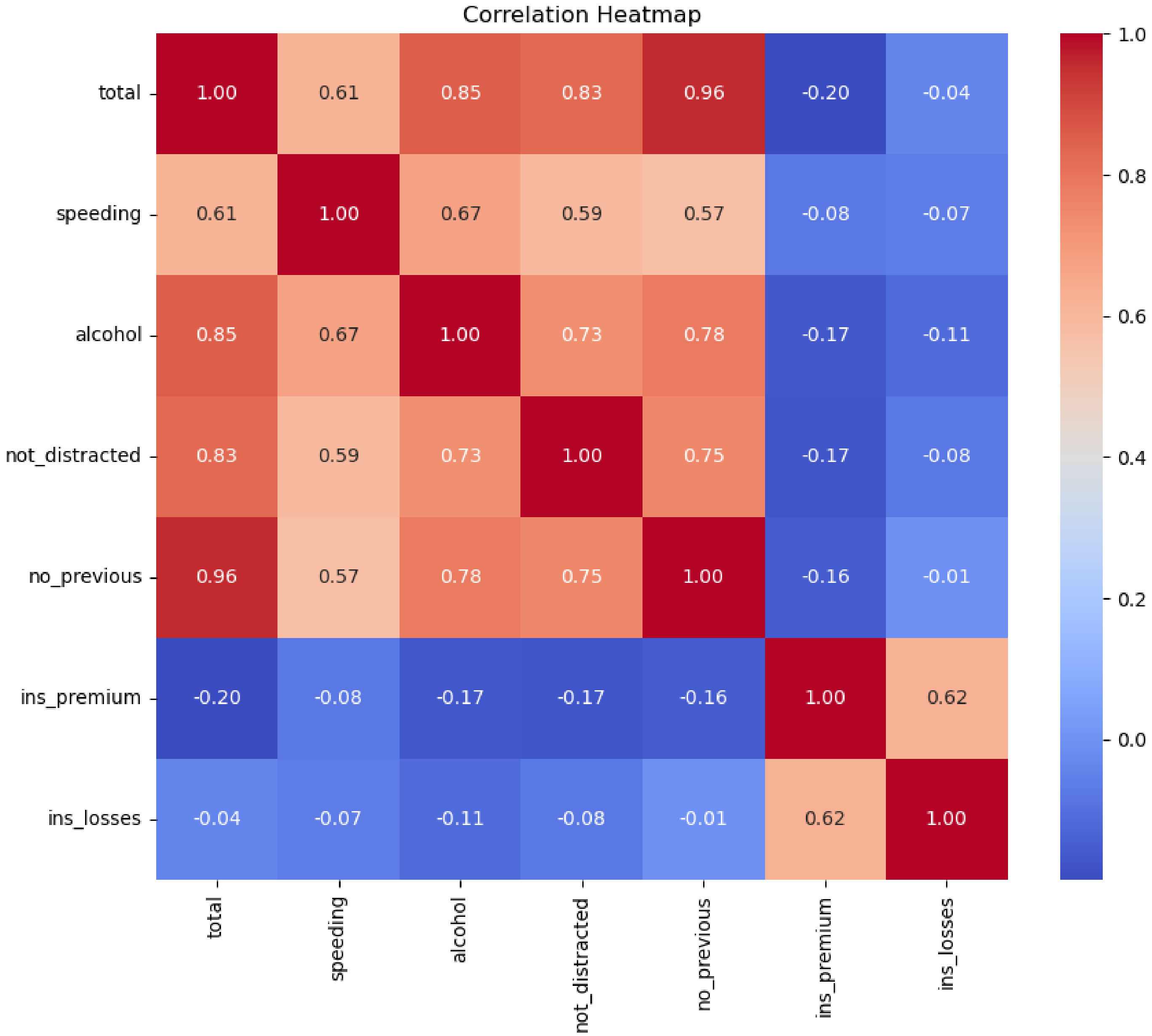
In [7]:

```
plt.figure(figsize=(10, 6))
sns.scatterplot(x="alcohol", y="speeding", data=crashes)
plt.xlabel("Alcohol-Related Crashes")
plt.ylabel("Speeding-Related Crashes")
plt.title("Scatter Plot of Alcohol vs. Speeding Crashes by State")
plt.show()
```



The scatter plot illustrates the relationship between alcohol-related car crashes and speeding-related car crashes by state. Each point represents a state, with its position on the graph indicating the number of incidents for both factors. The plot suggests that there is some positive correlation between alcohol-related and speeding-related crashes, indicating that states with higher alcohol-related crashes tend to also have higher speeding-related crashes.However,the strength and direction of this correlation may
vary across states

In [8]:
```
numeric_columns = ['total', 'speeding', 'alcohol', 'not_distracted', 'no_previous', 'ins
correlation_matrix = crashes[numeric_columns].corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```



The correlation heatmap indicates a moderately positive relationship between "alcohol-related crashes" and "insured losses," suggesting states with more alcohol-related crashes tend to have higher insured losses. Conversely, there's a negative correlation between "insured premiums" and "not distracted," implying states with fewer distracted driving incidents may have higher insurance premiums.

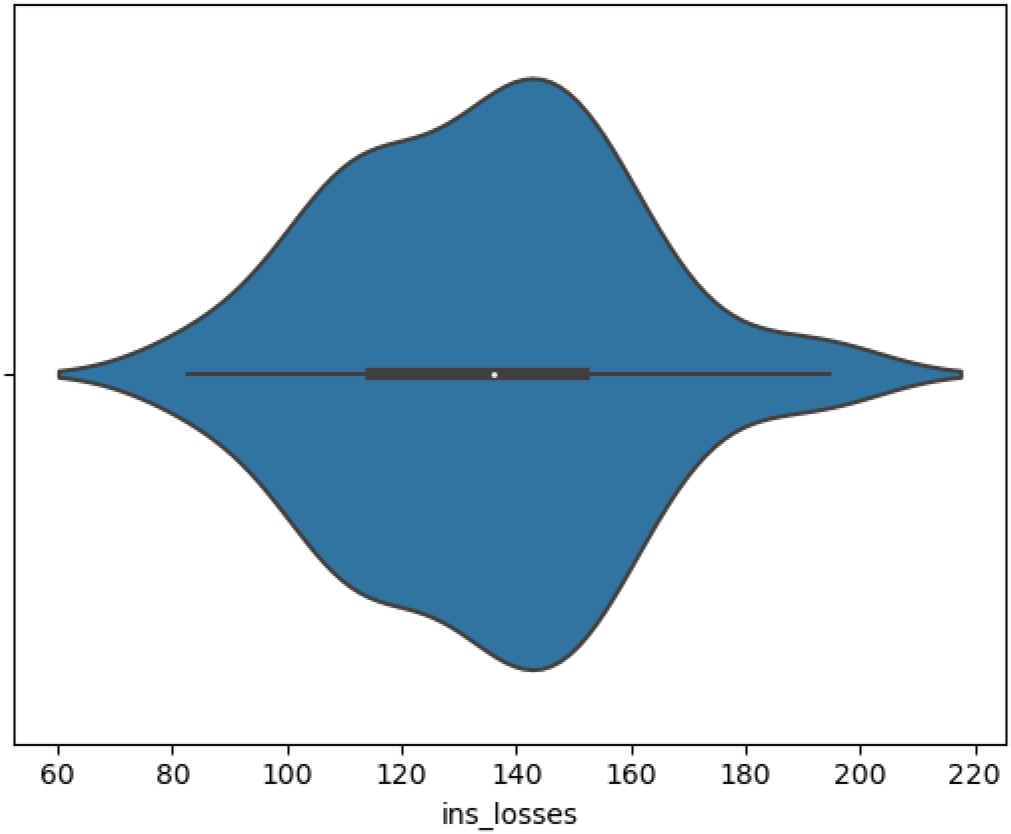
In [10]:

```
sns.violinplot(data=crashes, x="ins_losses")
```

Out[10]:

```
<Axes: xlabel='ins_losses'>
```

Loading [MathJax]/extensions/Safe.js



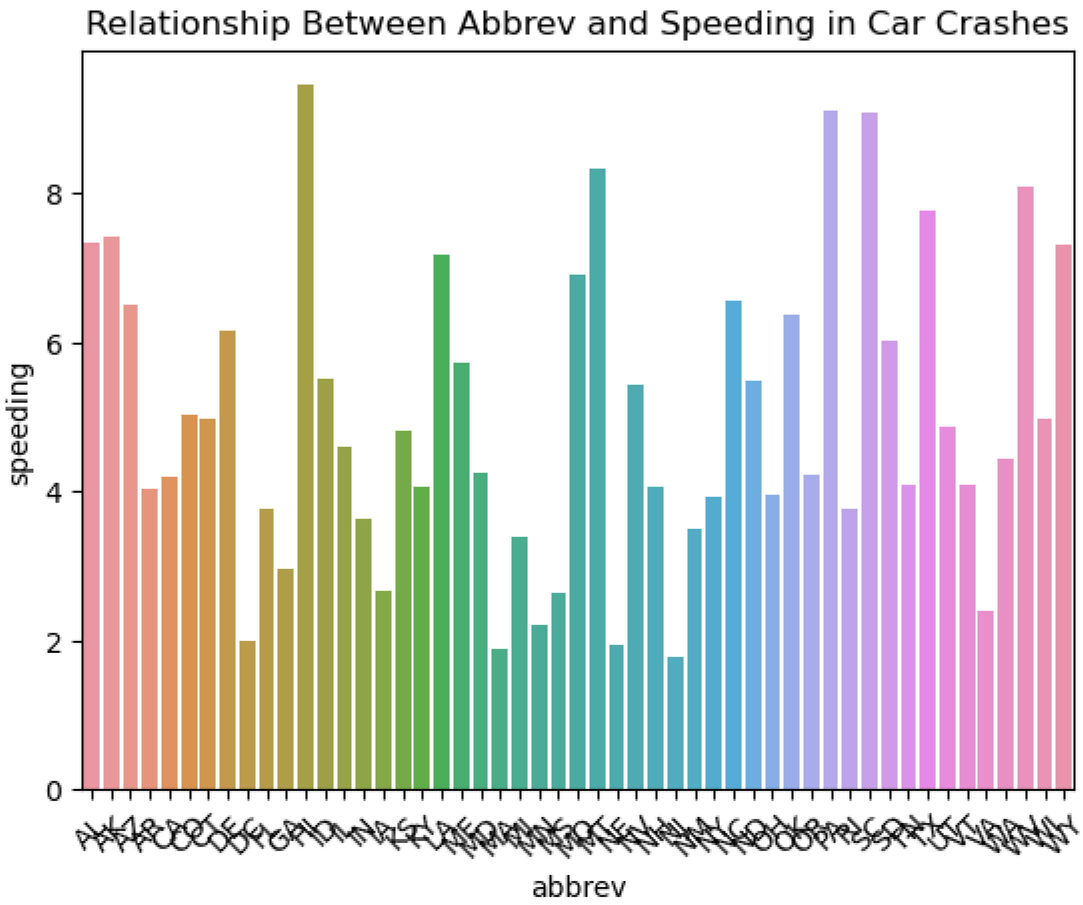
The violin plot illustrates the distribution of insured losses across different states.It reveals that the majority of states have relatively low insured losses, with a few outliers showing significantly higher losses. This suggests that most states experience relatively lower insurance losses, but a handful face more substantial financial losses due to car crashes.

In [12]:

```
sns.barplot(data=crashes, x="abbrev", y="speeding")
plt.xticks(rotation=45)
plt.title("Relationship Between Abbrev and Speeding in Car Crashes")
```

Out[12]:

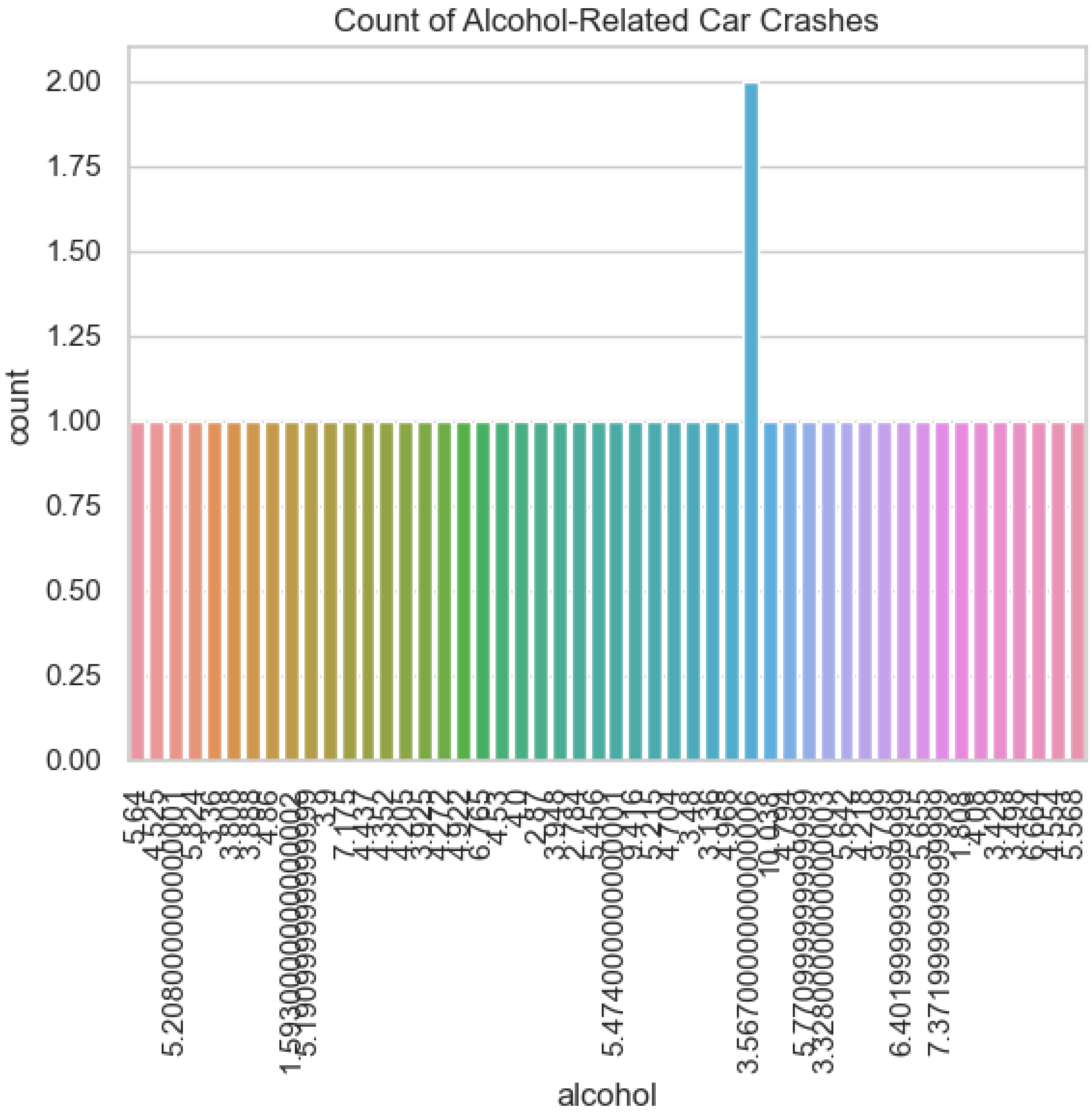
```
Text(0.5, 1.0, 'Relationship Between Abbrev and Speeding in Car Crashes')
```



The bar plot shows the relationship between state abbreviations (abbrev) and the frequency of speeding-related car crashes.It highlights variations in speeding incidents across different states. States with higher bars indicate a greater prevalence of speeding-related crashes, while those with lower bars have fewer such incidents. This suggests that there are regional differences in speeding-related car crash rates among the states represented.

In [14]:

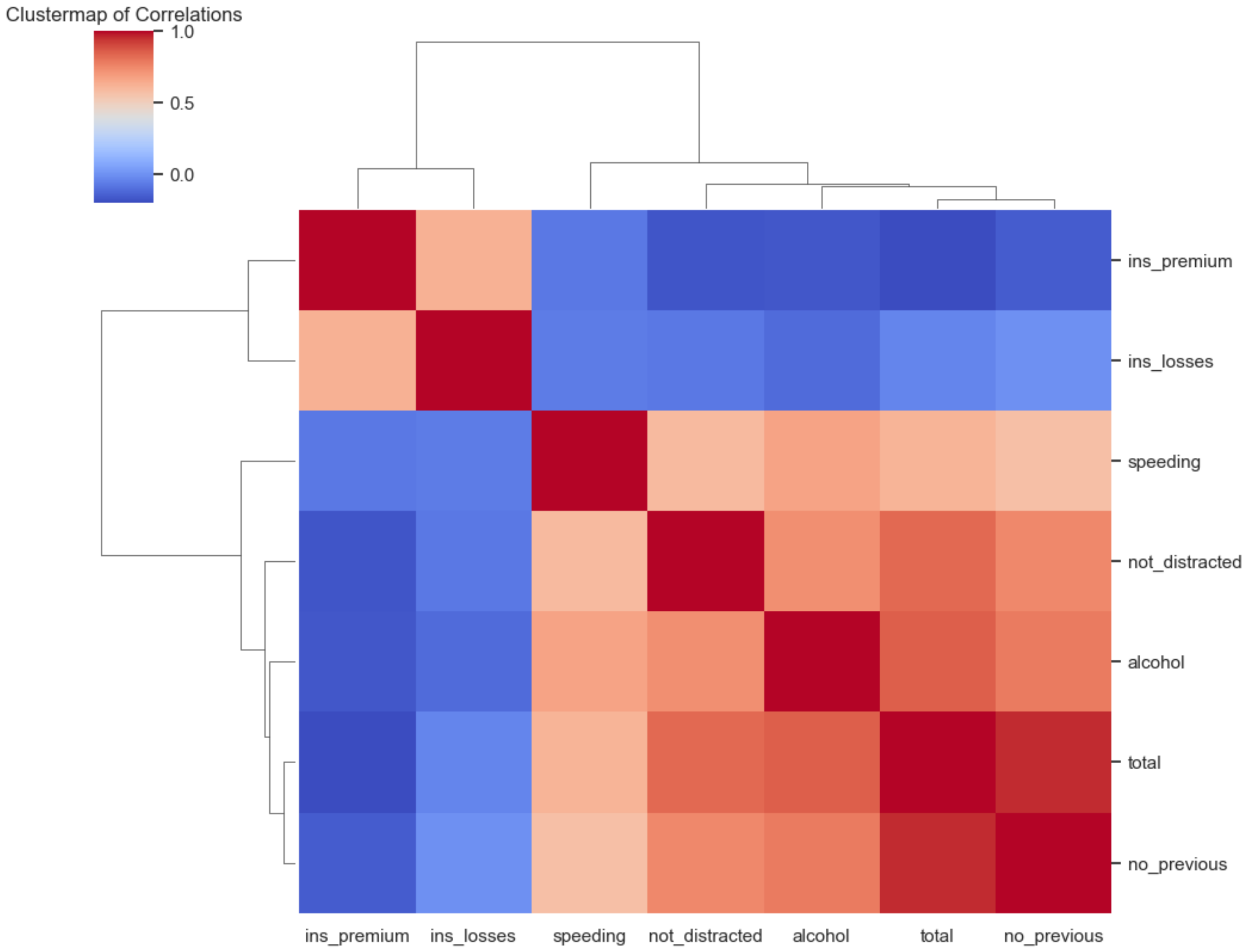
```
sns.set_theme(style="whitegrid")
selected_values =crashes['alcohol'].unique()[:50]
sns.countplot(x=crashes["alcohol"])
plt.xticks(range(len(selected_values)), selected_values, rotation=90)
plt.title("Count of Alcohol-Related Car Crashes")
plt.show()
```



The count plot displays the distribution of alcohol-related car crashes for the first 50 unique alcohol values in the dataset. It indicates that alcohol-related car crashes vary across these specific alcohol levels. The highest frequency of crashes appearsto occur at certain alcohol levels, while others have significantly fewer incidents, suggesting potential thresholds or patterns in alcohol-related accidents.

In [15]:

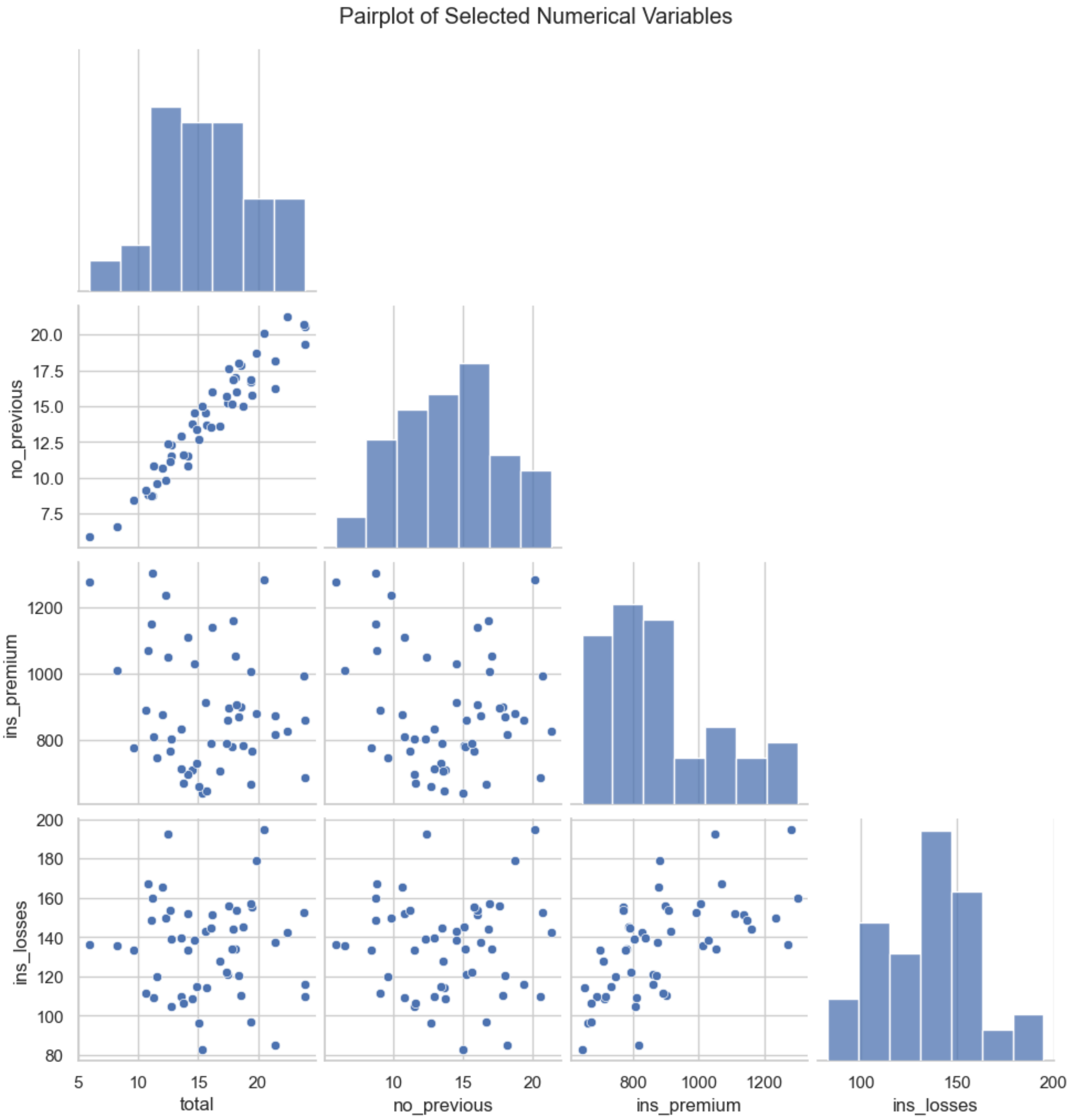
```
columns_for_cluster = crashes[['total', 'speeding', 'alcohol', 'not_distracted', 'no_pre
sns.clustermap(columns_for_cluster.corr(), cmap='coolwarm', figsize=(10, 8))
plt.title('Clustermap of Correlations')
plt.show()
```



The clustermap of correlations for selected variables in the "car_crashes" dataset reveals clusters of variables with similar correlations. Notably, "alcohol-related crashes" and "insured losses" are clustered together, indicating a positive correlation between these two factors. Conversely, "insured premiums" and "not distracted" are clustered, suggesting a negative correlation. This visualization helps identify groups of related variables and their potential impact on car crash statistics.

In [17]:

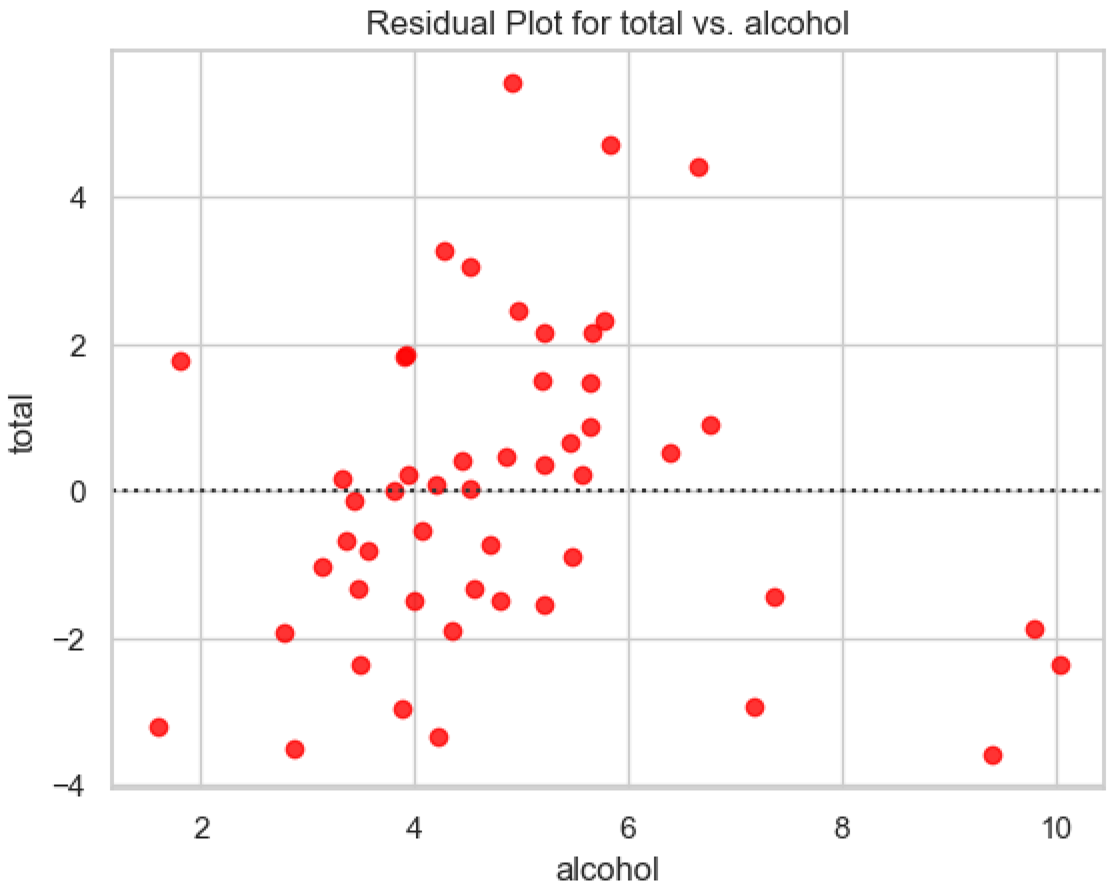
```
numerical_columns = ['total','no_previous', 'ins_premium', 'ins_losses']
sns.pairplot(crashes[numerical_columns],corner=True)
plt.suptitle("Pairplot of Selected Numerical Variables", y=1.02)
plt.show()
```



The pairplot of selected numerical variables from the "car_crashes" dataset shows scatterplots for each pair of variables, along with histograms on the diagonal. It provides insights into the relationships and distributions between these variables. From the pairplot, we can observe that there is a positive correlation between "insured losses" and "insurance premiums," and there is also a positive correlation between "total crashes" and "no previous crashes," indicating that states with higher total crashes tend to have higher numbers of crashes with no previous incidents.

In [20]:

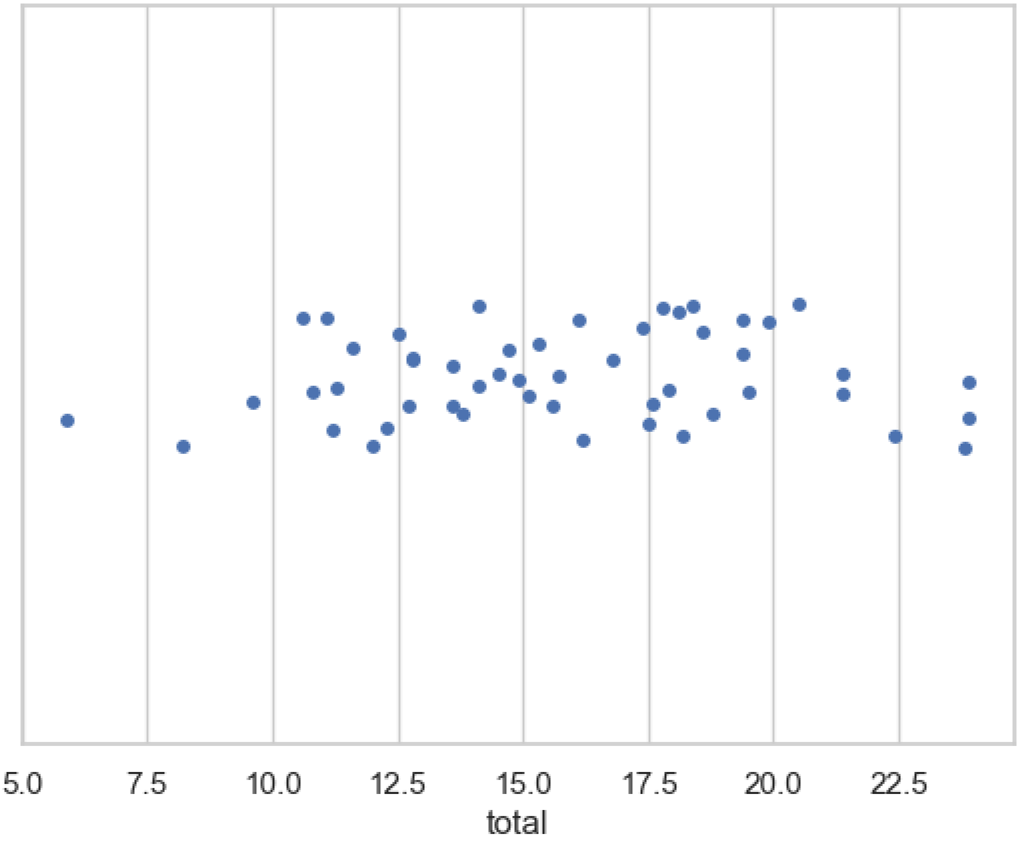
```
sns.residplot(x='alcohol', y='total', data=crashes, color='red')
plt.title(f"Residual Plot for {'total'} vs. {'alcohol'}")
plt.show()
```



The residual plot for "total crashes" versus "alcohol-related crashes" from the "car_crashes" dataset is used to assess the goodness of fit for a regression model. In this case, the red dots represent the residuals (differences between observed and predicted total crashes) for different alcohol-related crash levels. The plot indicates that the model may not capture all the variance in the data, as there are patterns in the residuals. For instance, it shows a slight U-shaped pattern, suggesting that the relationship between alcohol-related crashes and total crashes may not be perfectly linear, and there may be other factors influencing the total crash count.

```
In [21]: sns.stripplot(data=crashes, x="total")
```

```
Out[21]: <Axes: xlabel='total'>
```



The strip plot displays the distribution of "total crashes" from the "car_crashes" dataset along the x-axis. Each point represents a data point, and the plot helps visualize the frequency and distribution of total crashes. From the plot, we can observe that most states have a relatively low number of total crashes, with a few outliers indicating states with significantly higher crash counts.

In [ ]:

Loading [MathJax]/extensions/Safe.js