In [3]:

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

In [4]:

```
sns.get_dataset_names()
```

Out[4]:

```
['anagrams',
 'anscombe',
 'attention',
 'brain_networks',
 'car_crashes',
 'diamonds',
 'dots',
 'dowjones',
 'exercise',
 'flights',
 'fmri',
 'geyser',
 'glue',
 'healthexp',
 'iris',
 'mpg',
 'penguins',
 'planets',
 'seaice',
 'taxis',
 'tips',
 'titanic']
```

In [6]:

```
dataset=sns.load_dataset('car_crashes')
```

In [7]:

```
dataset.head()
```

Out[7]:

	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 [8]:

```
dataset.tail()
```

Out[8]:

	total	speeding	alcohol	not_distracted	no_previous	ins_premium	ins_losses	abbrev
46	12.7	2.413	3.429	11.049	11.176	768.95	153.72	VA
47	10.6	4.452	3.498	8.692	9.116	890.03	111.62	WA
48	23.8	8.092	6.664	23.086	20.706	992.61	152.56	WV
49	13.8	4.968	4.554	5.382	11.592	670.31	106.62	WI
50	17.4	7.308	5.568	14.094	15.660	791.14	122.04	WY

In [9]:

```
#describing dataset
dataset.describe()
```

Out[9]:

	total	speeding	alcohol	not_distracted	no_previous	ins_premium	ins_loss
count	51.000000	51.000000	51.000000	51.000000	51.000000	51.000000	51.0000
mean	15.790196	4.998196	4.886784	13.573176	14.004882	886.957647	134.4931
std	4.122002	2.017747	1.729133	4.508977	3.764672	178.296285	24.8359
min	5.900000	1.792000	1.593000	1.760000	5.900000	641.960000	82.7500
25%	12.750000	3.766500	3.894000	10.478000	11.348000	768.430000	114.6450
50%	15.600000	4.608000	4.554000	13.857000	13.775000	858.970000	136.0500
75%	18.500000	6.439000	5.604000	16.140000	16.755000	1007.945000	151.8700
max	23.900000	9.450000	10.038000	23.661000	21.280000	1301.520000	194.7800

Univariate

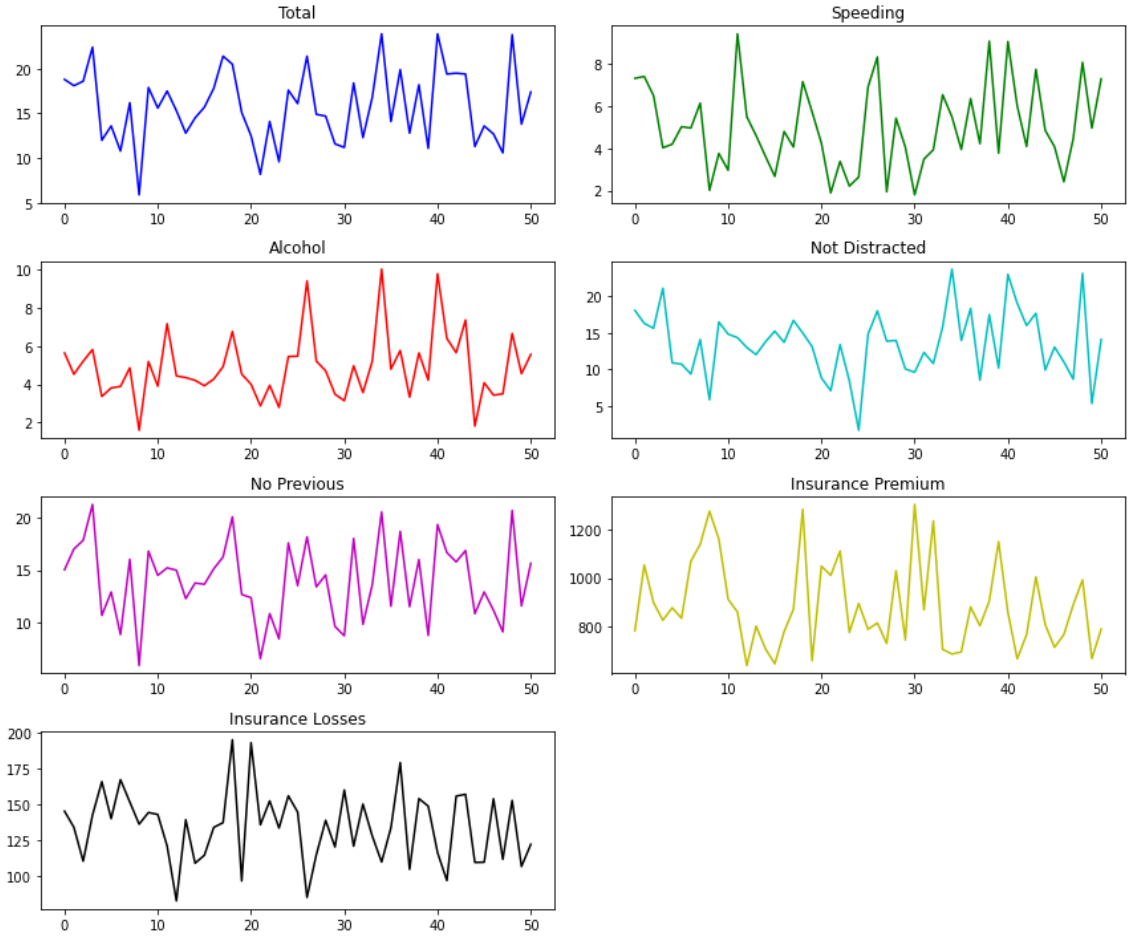
Definition and Objective:

Univariate is a statistical term that refers to the analysis or study of a single variable or data set at a time. It involves examining and describing the characteristics, distribution, and patterns of variation within a single set of data, typically represented by a single column or variable in a dataset. Univariate analysis is often used to summarize and gain insights into the behavior of individual variables without considering their relationships with other variables.

In [14]:

```
plt.figure(figsize=(12, 10))

plt.subplot(4, 2, 1)
plt.plot(dataset['total'], 'b')
plt.title('Total')

"""
Total (Blue Line):
The graph shows the trend in total car crashes over the dataset.
Inference: The overall number of auto accidents varies substantially over time, but no c
"""


plt.subplot(4, 2, 2)
plt.plot(dataset['speeding'], 'g')
plt.title('Speeding')

"""
Speeding (Green Line):
This graph represents the trend in car crashes caused by speeding.
Inference: The number of speed-related car accidents seems to fluctuate, but it doesn't

"""

plt.subplot(4, 2, 3)
plt.plot(dataset['alcohol'], 'r')
plt.title('Alcohol')

"""
Alcohol (Red Line):
The graph displays the trend in car crashes related to alcohol consumption.
Inference: Although there is substantial fluctuation in alcohol-related car accidents, t

"""


plt.subplot(4, 2, 4)
plt.plot(dataset['not_distracted'], 'c')
plt.title('Not Distracted')

"""
Not Distracted (Cyan Line):
This graph illustrates the trend in car crashes where drivers were not distracted.
Inference: The number of car crashes by non-distracted drivers shows fluctuations, but n
"""

plt.subplot(4, 2, 5)
plt.plot(dataset['no_previous'], 'm')
plt.title('No Previous')

"""
No Previous (Magenta Line):
The graph shows the trend in car crashes by drivers with no previous incidents.
Inference: There are minor fluctuations but no clear trend in car accidents involving dr

"""


plt.subplot(4, 2, 6)
plt.plot(dataset['ins_premium'], 'y')
plt.title('Insurance Premium')
```

```
"""
Insurance Premium (Yellow Line):
This graph represents the trend in insurance premiums.
Inference: The graph, which appears to vary without following a defined pattern, doesn't

"""

plt.subplot(4, 2, 7)
plt.plot(dataset['ins_losses'], 'k')
plt.title('Insurance Losses')

"""
Insurance Losses (Black Line):
The graph displays the trend in insurance losses.
Inference: Insurance losses seem to change randomly, much like insurance rates do.

"""

plt.tight_layout()
```

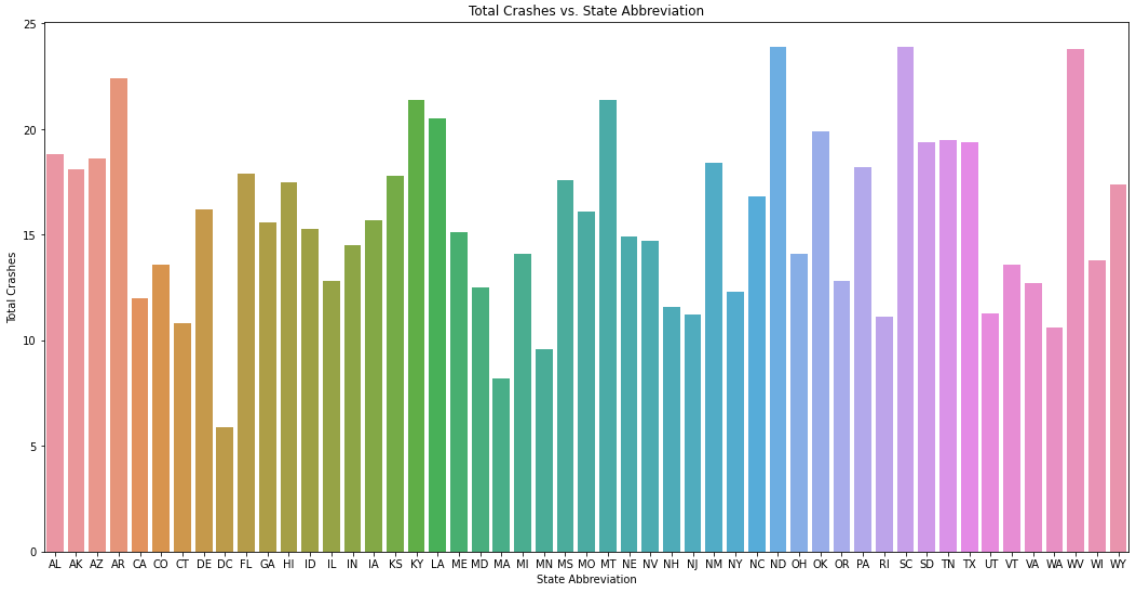


In [16]:

```
plt.figure(figsize=(18, 9))
sns.barplot(data=dataset,x='abbrev', y='total')
plt.xlabel('State Abbreviation')
plt.ylabel('Total Crashes')
plt.title('Total Crashes vs. State Abbreviation')
```

Out[16]:

```
Text(0.5, 1.0, 'Total Crashes vs. State Abbreviation')
```

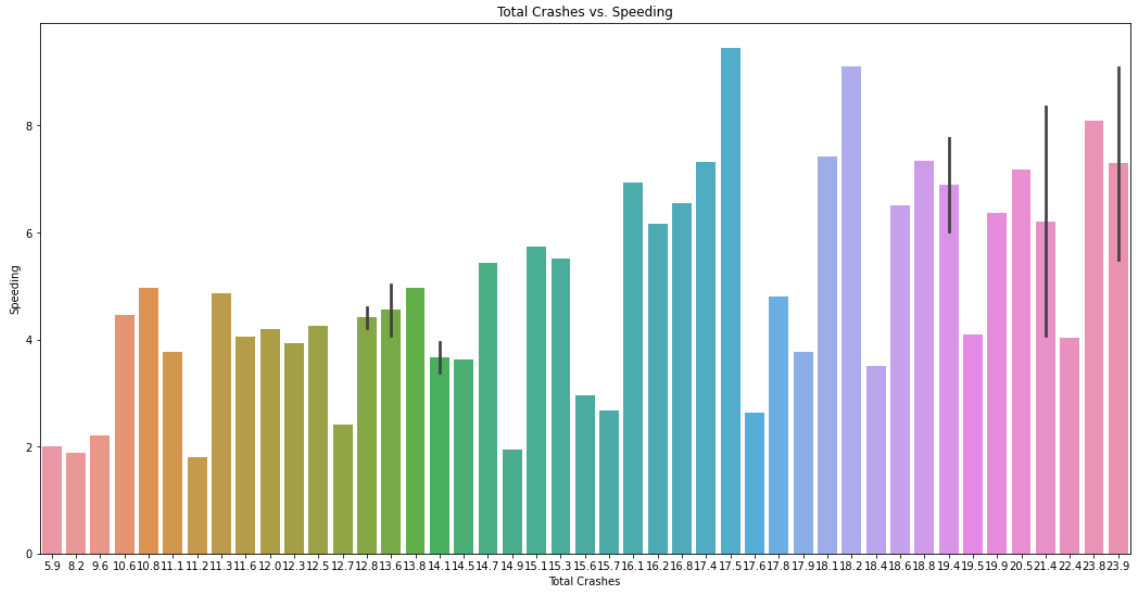


In [38]:

```
plt.figure(figsize=(18, 9))
sns.barplot(data=dataset,x='total', y='speeding')
plt.ylabel('Speeding')
plt.xlabel('Total Crashes')
plt.title('Total Crashes vs. Speeding')

"""
Inference:
The x-axis shows the overall number of collisions, and the y-axis shows how many of thos
The figure enables us to analyse how speeding affects the total number of auto accidents
There is a general tendency of an increase in crashes involving speeding as the overall
This shows that the fraction of crashes involving speeding tends to increase along with
Understanding the effects of speeding on overall road safety and developing specific str

"""
```

Out[38]:

```
'\nInference:\nThe x-axis shows the overall number of collisions, and the
y-axis shows how many of those collisions involved speeding.\nThe figure
enables us to analyse how speeding affects the total number of auto accid
ents.\nThere is a general tendency of an increase in crashes involving sp
eeding as the overall number of crashes rises.\nThis shows that the fract
ion of crashes involving speeding tends to increase along with the overal
l number of car crashes.\nUnderstanding the effects of speeding on overal
l road safety and developing specific strategies to lower accidents cause
d by speeding can both benefit from an analysis of this relationship.\n
\n'
```



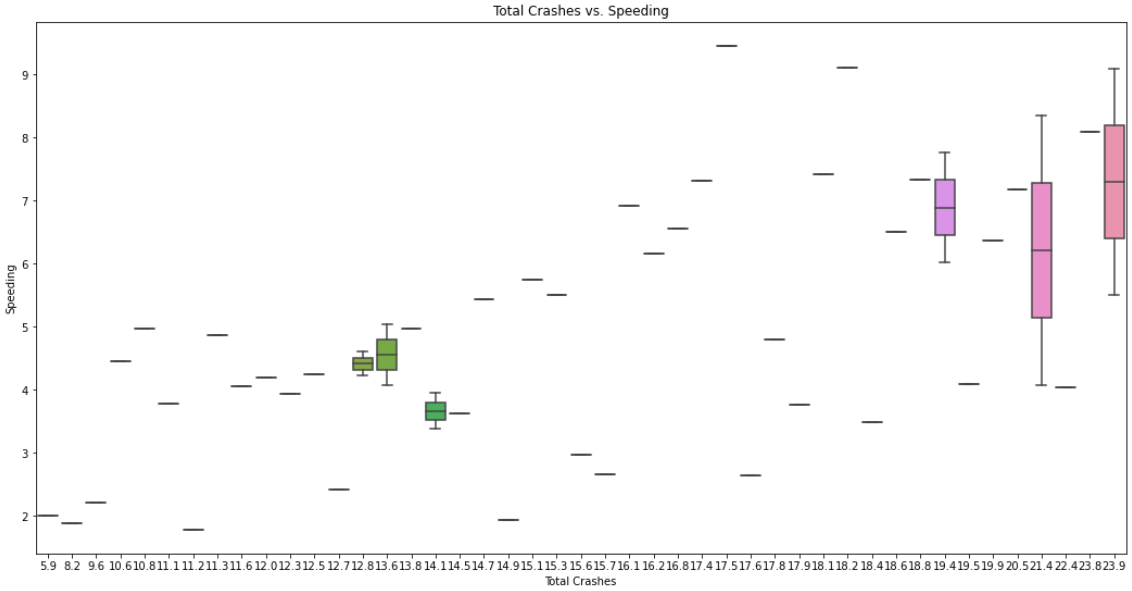
Boxplot

In [19]:

```
plt.figure(figsize=(18,9))
sns.boxplot(x="total",y="speeding",data=dataset)
plt.ylabel('Speeding')
plt.xlabel('Total Crashes')
plt.title('Total Crashes vs. Speeding')
```

Out[19]:

```
Text(0.5, 1.0, 'Total Crashes vs. Speeding')
```

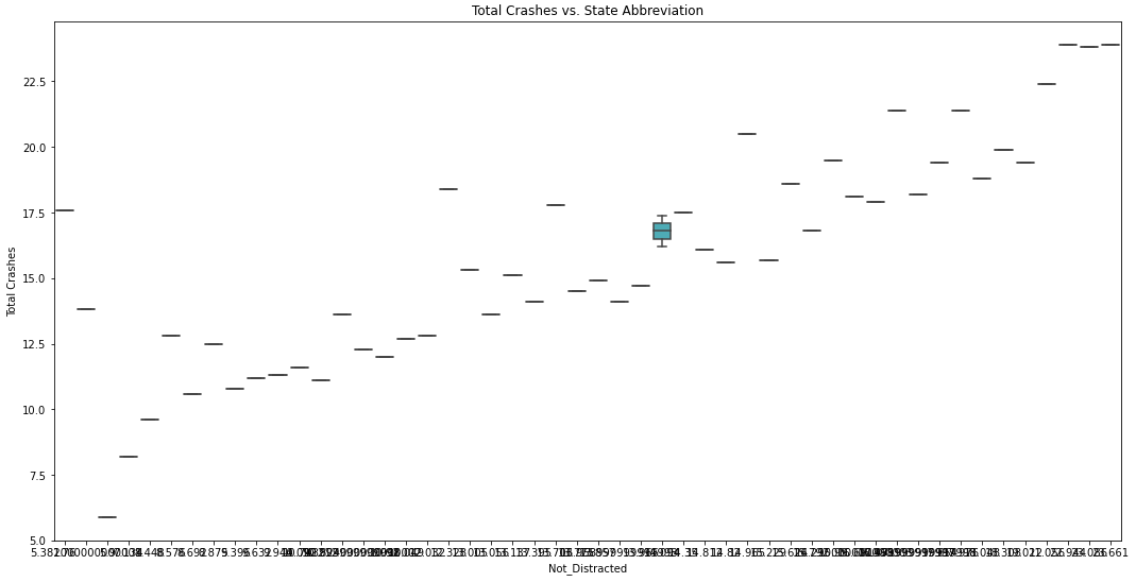


In [39]:

```
plt.figure(figsize=(18,9))
sns.boxplot(x="not_distracted",y="total",data=dataset)
plt.xlabel('Not_Distracted')
plt.ylabel('Total Crashes')
plt.title('Total Crashes vs. State Abbreviation')

"""
Inference :
The box plot depicts the distribution of total crashes according to the level of driver
It offers information on how distraction impacts the overall number of auto accidents.
According to the level of driving distraction, the overall number of crashes varies, pot
This raises the possibility that non-distracted drivers may be involved in more collisic

"""
```

Out[39]:

```
'\nInference :\nThe box plot depicts the distribution of total crashes ac
cording to the level of driver distraction (Not Distracted).\nIt offers i
nformation on how distraction impacts the overall number of auto accident
s.\nAccording to the level of driving distraction, the overall number of
crashes varies, potentially increasing when drivers are not distracted.\n
This raises the possibility that non-distracted drivers may be involved i
n more collisions, highlighting the necessity of researching the factors
that contribute to distraction and unsafe driving practises in order to i
ncrease road safety.\n\n'
```



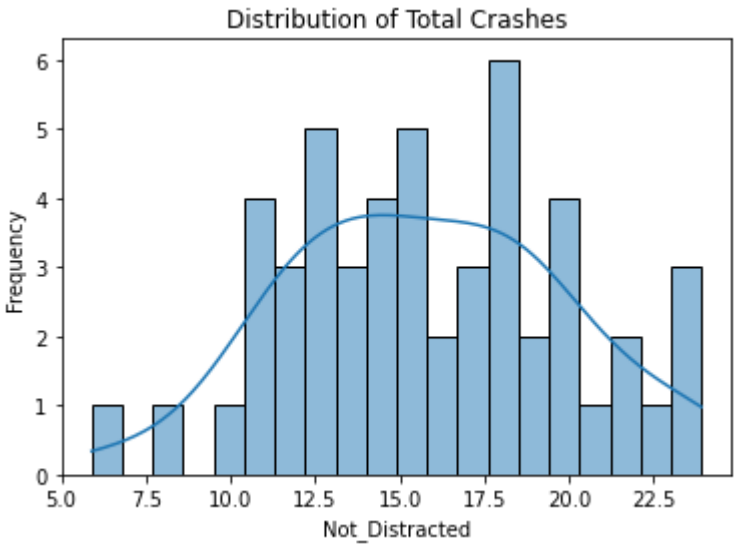
Histogram

In [40]:

```
sns.histplot(data=dataset, x='total', bins=20, kde=True)
plt.xlabel('Not_Distracted')
plt.ylabel('Frequency')
plt.title('Distribution of Total Crashes')

"""
Inference :
The distribution of all auto accidents is shown in the histogram.n.
The plot demonstrates that most observations fall into a narrow range of total crashes,
The distribution is right-skewed, which means that some instances have considerably grea
Understanding the distribution of all crashes through this visualisation can assist find

"""
```

Out[40]:

```
'\nInference :\nThe distribution of all auto accidents is shown in the hi
stogram.n.\nThe plot demonstrates that most observations fall into a narr
ow range of total crashes, with a peak in frequency.\nThe distribution is
right-skewed, which means that some instances have considerably greater c
rash counts than others.\nUnderstanding the distribution of all crashes t
hrough this visualisation can assist find typical crash count ranges and
outliers in the dataset.\n\n'
```

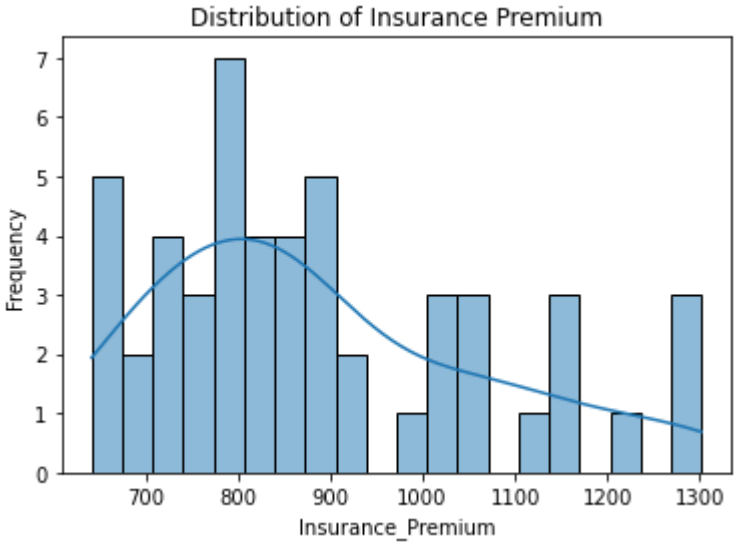


In [41]:

```
sns.histplot(data=dataset, x='ins_premium', bins=20, kde=True)
plt.xlabel('Insurance_Premium')
plt.ylabel('Frequency')
plt.title('Distribution of Insurance Premium')

"""
Inference :
The distribution of insurance premiums is shown by the histogram.
The plot demonstrates that the distribution's peaks are formed by the most prevalent ins
The right-skewed distribution suggests that certain observations have very high insuranc
This visualisation offers insights into typical premium ranges and probable outliers, as

"""
```

Out[41]:

```
"\nInference :\nThe distribution of insurance premiums is shown by the hi
stogram.\nThe plot demonstrates that the distribution's peaks are formed
by the most prevalent insurance premium ranges, which have greater freque
ncies.\nThe right-skewed distribution suggests that certain observations
have very high insurance premiums.\nThis visualisation offers insights in
to typical premium ranges and probable outliers, assisting in understandi
ng the distribution of insurance premiums within the dataset.\n\n"
```

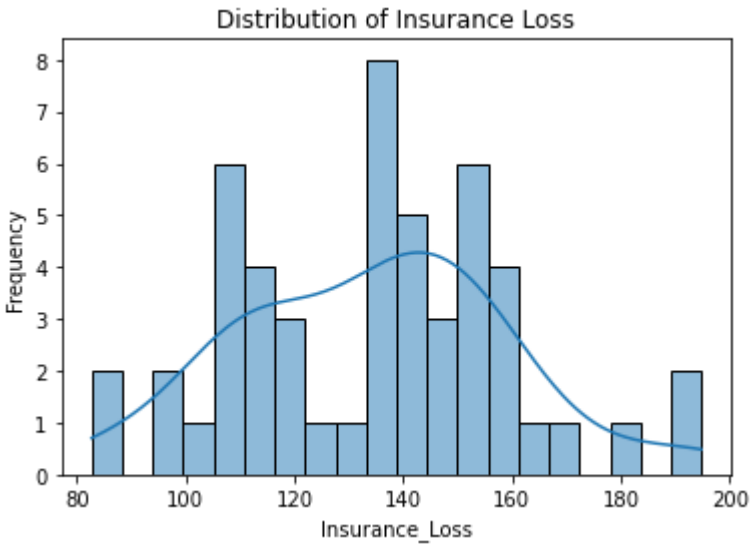


In [23]:

```
sns.histplot(data=dataset, x='ins_losses', bins=20, kde=True)
plt.xlabel('Insurance_Loss')
plt.ylabel('Frequency')
plt.title('Distribution of Insurance Loss')

"""
Inference :
The histogram represents the distribution of insurance losses.
The plot indicates that the majority of insurance losses fall within specific ranges, wi
The distribution appears right-skewed, indicating that a few instances have considerably
This visualization helps in understanding the distribution of insurance losses within th
"""
```

Out[23]:

```
'\nInference :\nThe histogram represents the distribution of insurance lo
sses.\nThe plot indicates that the majority of insurance losses fall with
in specific ranges, with peaks in frequency.\nThe distribution appears ri
ght-skewed, indicating that a few instances have considerably higher insu
rance losses.\nThis visualization helps in understanding the distribution
of insurance losses within the dataset, highlighting common loss ranges a
nd potential outliers.\n'
```



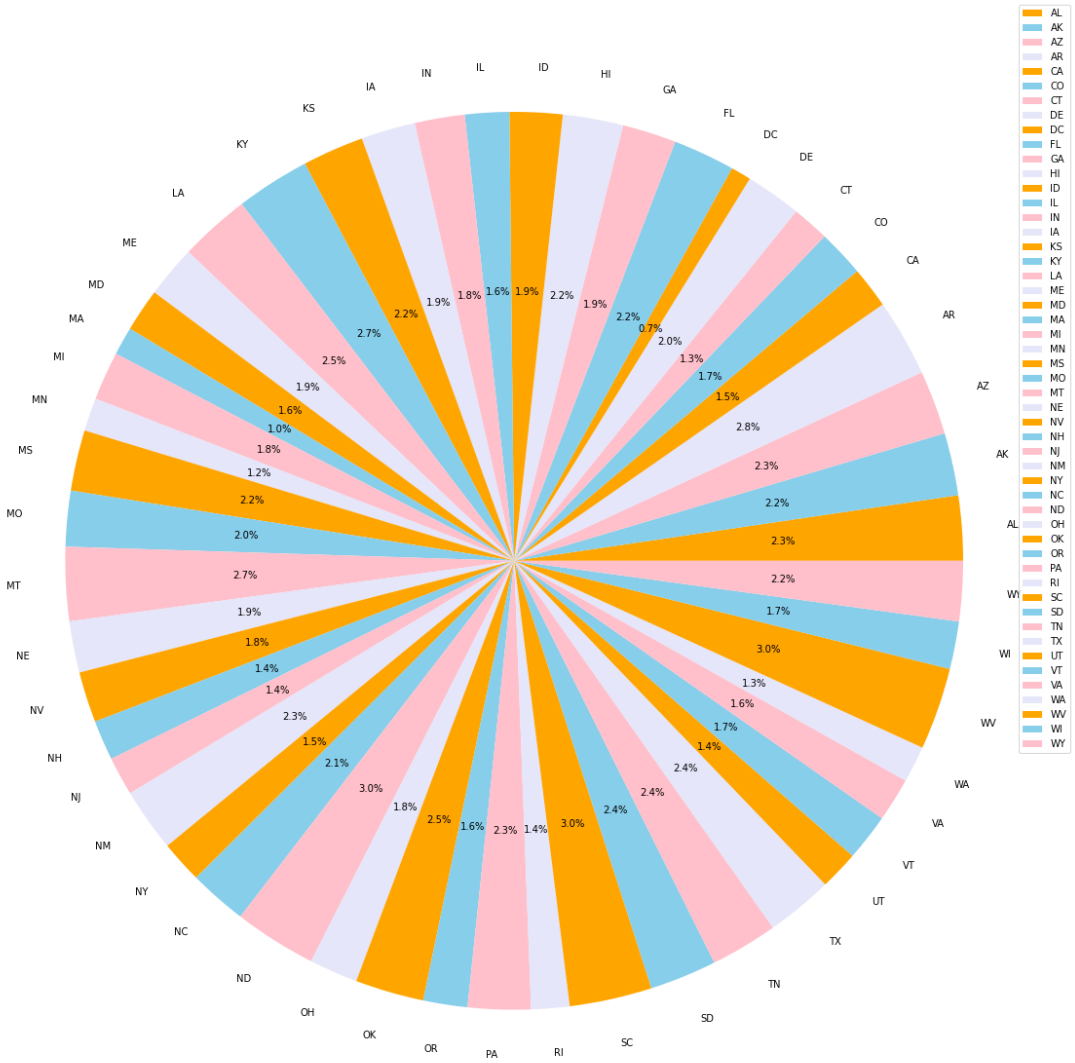
Piechart

In [42]:

```
fig = plt.figure(figsize=(20,20))
axes1 = fig.add_axes([0.1,0.1,0.8,0.8])  # (left,bottom,width,height)
axes1.pie(dataset['total'],labels=dataset['abbrev'],autopct='%0.1f%%',colors =['orange',
axes1.legend()

"""
Inference :
The state abbreviations on the pie chart serve to illustrate how many total auto acciden
Each piece of the pie symbolises a state, and the proportion of total crashes in each st
The state abbreviations are shown by the labels on the chart.
The state that each slice represents can be determined using the key in the legend.
This pie chart makes it easy to compare how much each state contributed to the total num

"""
```

Out[42]:

```
'\nInference :\nThe state abbreviations on the pie chart serve to illustr
ate how many total auto accidents occurred in each state.\nEach piece of
the pie symbolises a state, and the proportion of total crashes in each s
tate is shown by the size of the piece.\nThe state abbreviations are show
n by the labels on the chart.\nThe state that each slice represents can b
e determined using the key in the legend.\nThis pie chart makes it easy t
o compare how much each state contributed to the total number of automobi
le crashes in the dataset.\n\n'
```


AL
AK
AZ
AR
CA
CO
CT
DE
DC
FL
GA
HI
ID
IL
IN
IA
KS
KY
LA
ME
MD
MA
MI
MN
MS
MO
MT
NE
NV
NH
NJ
NM
NY
NC
ND
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VT
VA
WA
WV
WI
WY

Bivariate

Definition and Objective:

Bivariate is a statistical term that pertains to the analysis or study of the relationship between two variables simultaneously. In a bivariate analysis, the focus is on understanding how changes in one variable relate to or influence changes in another variable. This type of analysis is essential for examining associations, dependencies, correlations, and cause-and-effect relationships between two specific variables.

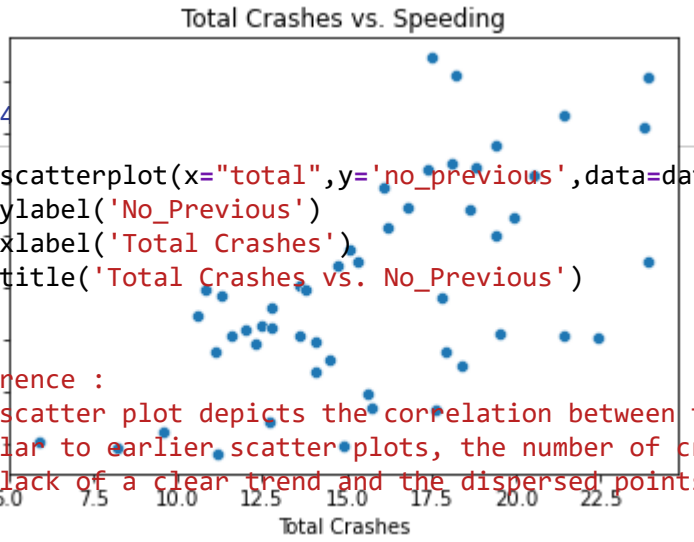
Scatterplot

In [43]:

```
sns.scatterplot(x="total",y='speeding',data=dataset)
plt.ylabel('Speeding')
plt.xlabel('Total Crashes')
plt.title('Total Crashes vs. Speeding')

"""
Inference :
The scatter figure illustrates the correlation between the overall number of auto accide
Based on this scatter plot, there doesn't seem to be a significant linear link between t
The lack of a discernible trend and the scattered placement of the points on the plot su
To accurately quantify the relationship between these factors, more statistical analysis

"""
```

Out[43]:

```
"\nInference :\nThe scatter figure illustrates the correlation between th
e overall number of auto accidents and those involving speeding.\nBased o
n this scatter plot, there doesn't seem to be a significant linear link b
etween total crashes and speeding events.\nThe lack of a discernible tren
d and the scattered placement of the points on the plot suggest that ther
e may not be a strong correlation between total crashes and speeding.\nTo
accurately quantify the relationship between these factors, more statisti
cal analysis may be required.\n\n"
```

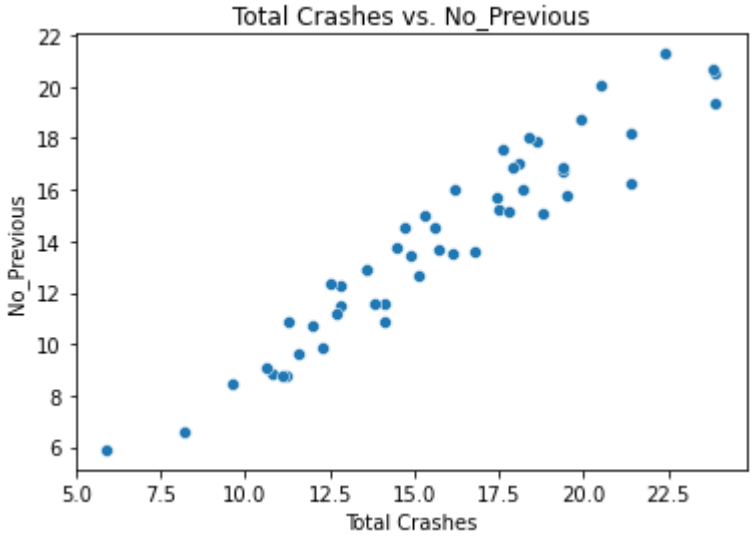


In [4

```
sns.scatterplot(x="total",y='no_previous',data=dataset)
plt.ylabel('No_Previous')
plt.xlabel('Total Crashes')
plt.title('Total Crashes vs. No_Previous')

"""
Inference :
The scatter plot depicts the correlation between the overall number of auto accidents an
Similar to earlier scatter plots, the number of crashes and crashes involving drivers wi
The lack of a clear trend and the dispersed points indicate that the number of crashes m
"""
```

Out[44]:

```
"\nInference :\nThe scatter plot depicts the correlation between the over
all number of auto accidents and accidents involving drivers who had no p
rior violations.\nSimilar to earlier scatter plots, the number of crashes
and crashes involving drivers with no prior incidents don't have a clear
linear link.\nThe lack of a clear trend and the dispersed points indicate
that the number of crashes may not be directly related to drivers' lack o
f prior occurrences. Perhaps more research is required.\n\n"
```



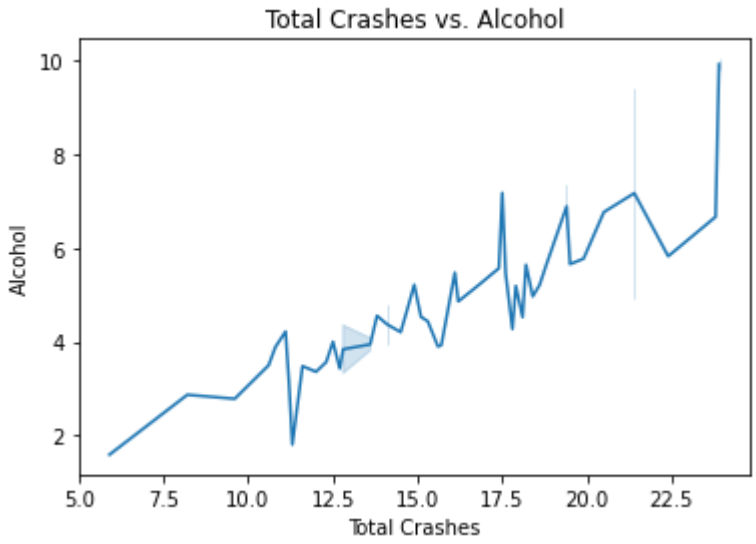
Lineplot

In [45]:

```
sns.lineplot(x="total",y="alcohol",data=dataset)
plt.ylabel('Alcohol')
plt.xlabel('Total Crashes')
plt.title('Total Crashes vs. Alcohol')
"""
Inference :
The link between overall auto accidents and drunk driving accidents is depicted by the l
It illustrates the variation in alcohol-related crashes relative to overall crash rates.
The points on the line are dispersed and lack a discernible pattern, hence there is no c
This shows that there might not be a clear association between the overall number of cra

"""
```

Out[45]:

```
'\nInference :\nThe link between overall auto accidents and drunk driving
accidents is depicted by the line plot.\nIt illustrates the variation in
alcohol-related crashes relative to overall crash rates.\nThe points on t
he line are dispersed and lack a discernible pattern, hence there is no o
bvious linear link.\nThis shows that there might not be a clear associati
on between the overall number of crashes and occurrences involving alcoho
l, necessitating further investigation.\n\n'
```

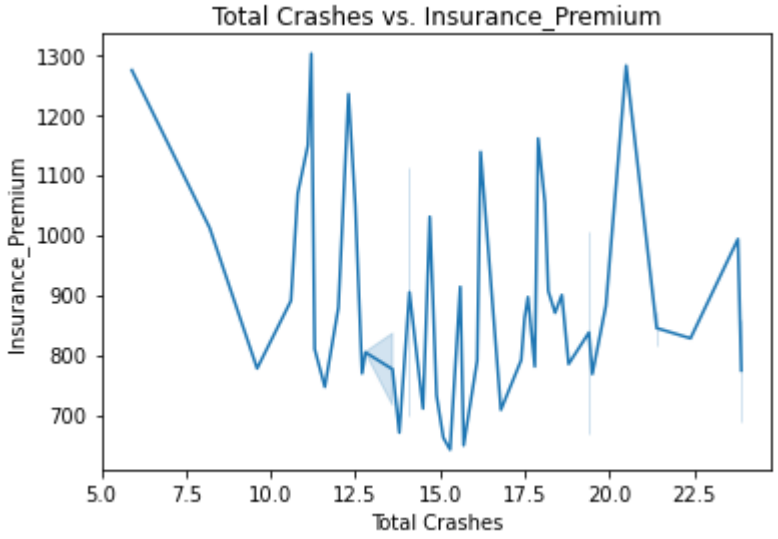


In [46]:

```
sns.lineplot(x="total",y="ins_premium",data=dataset)
plt.ylabel('Insurance_Premium')
plt.xlabel('Total Crashes')
plt.title('Total Crashes vs. Insurance_Premium')

"""
Inference :
The association between the total number of auto accidents and insurance rates is shown
It illustrates the relationship between insurance costs and the overall number of collis
The line's points are strewn across the map without any discernible pattern, and there i
Further research is required since this shows that there may not be a direct association

"""
```

Out[46]:

```
"\nInference :\nThe association between the total number of auto accident
s and insurance rates is shown by the line plot.\nIt illustrates the rela
tionship between insurance costs and the overall number of collisions.\nT
he line's points are strewn across the map without any discernible patter
n, and there is no obvious linear trend.\nFurther research is required si
nce this shows that there may not be a direct association between the tot
al number of crashes and insurance premiums.\n\n"
```



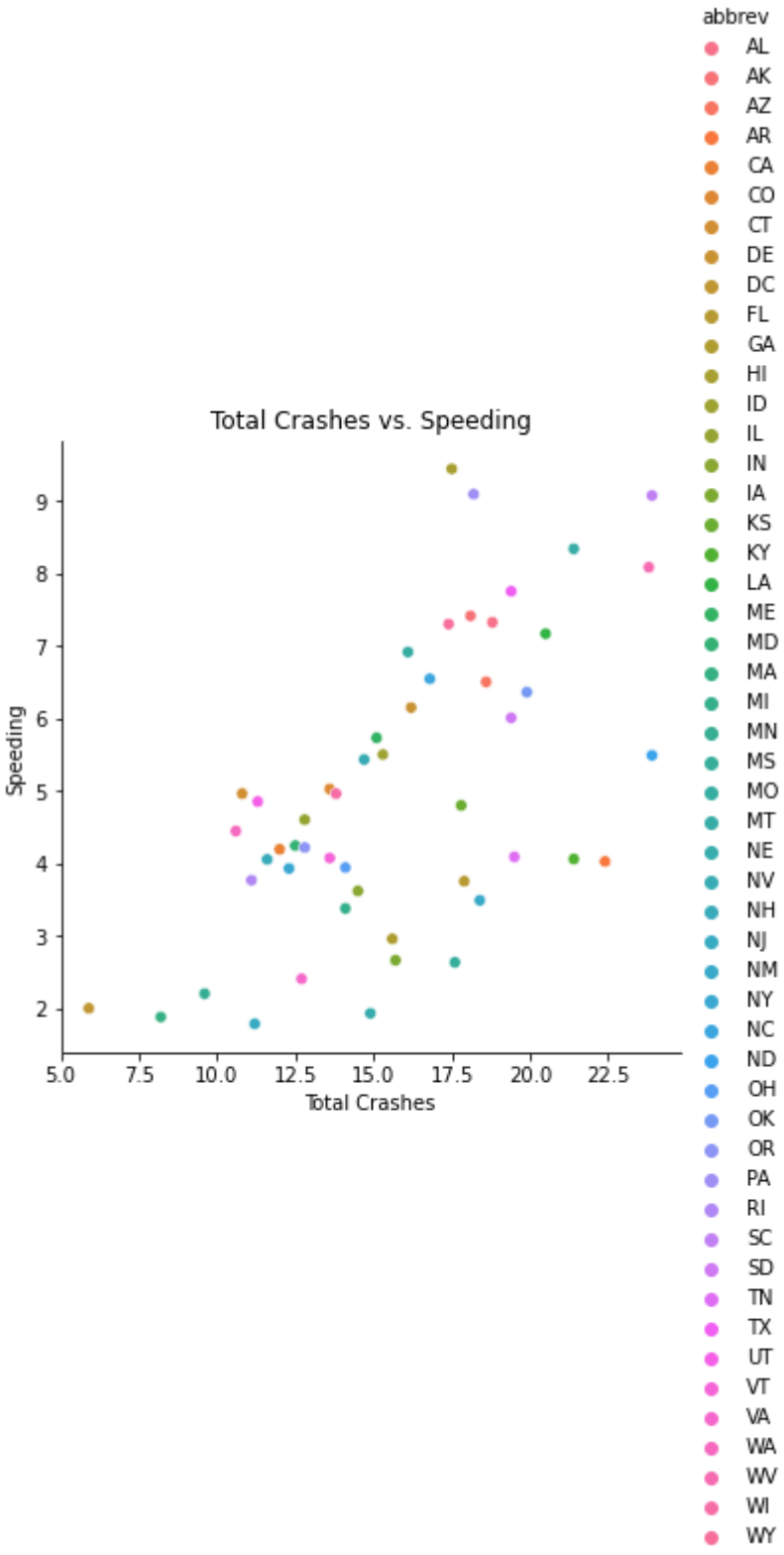
Replot

In [47]:

```
sns.relplot(x="total",y="speeding",data=dataset,hue="abbrev")
plt.ylabel('Speeding')
plt.xlabel('Total Crashes')
plt.title('Total Crashes vs. Speeding')

"""
Inference :
The association between overall auto accidents and accidents involving speeding is shown
Each point in the collection represents a data point, with various states indicated by d
The figure makes it easy to quickly see how the total number of crashes in various state
The lack of a defined pattern and apparent linear trend suggest that the relationship be

"""
```

Out[47]:

```
'\nInference :\nThe association between overall auto accidents and accide
nts involving speeding is shown by the relational plot ("relplot").\nEach
point in the collection represents a data point, with various states indi
cated by distinct colours (hue).\nThe figure makes it easy to quickly see
how the total number of crashes in various states varies when speeding is
a contributing factor.\nThe lack of a defined pattern and apparent linear
trend suggest that the relationship between total collisions and speeding
incidents may not be simple and may differ by state. To investigate state
-specific trends, additional study could be needed.\n\n'
```

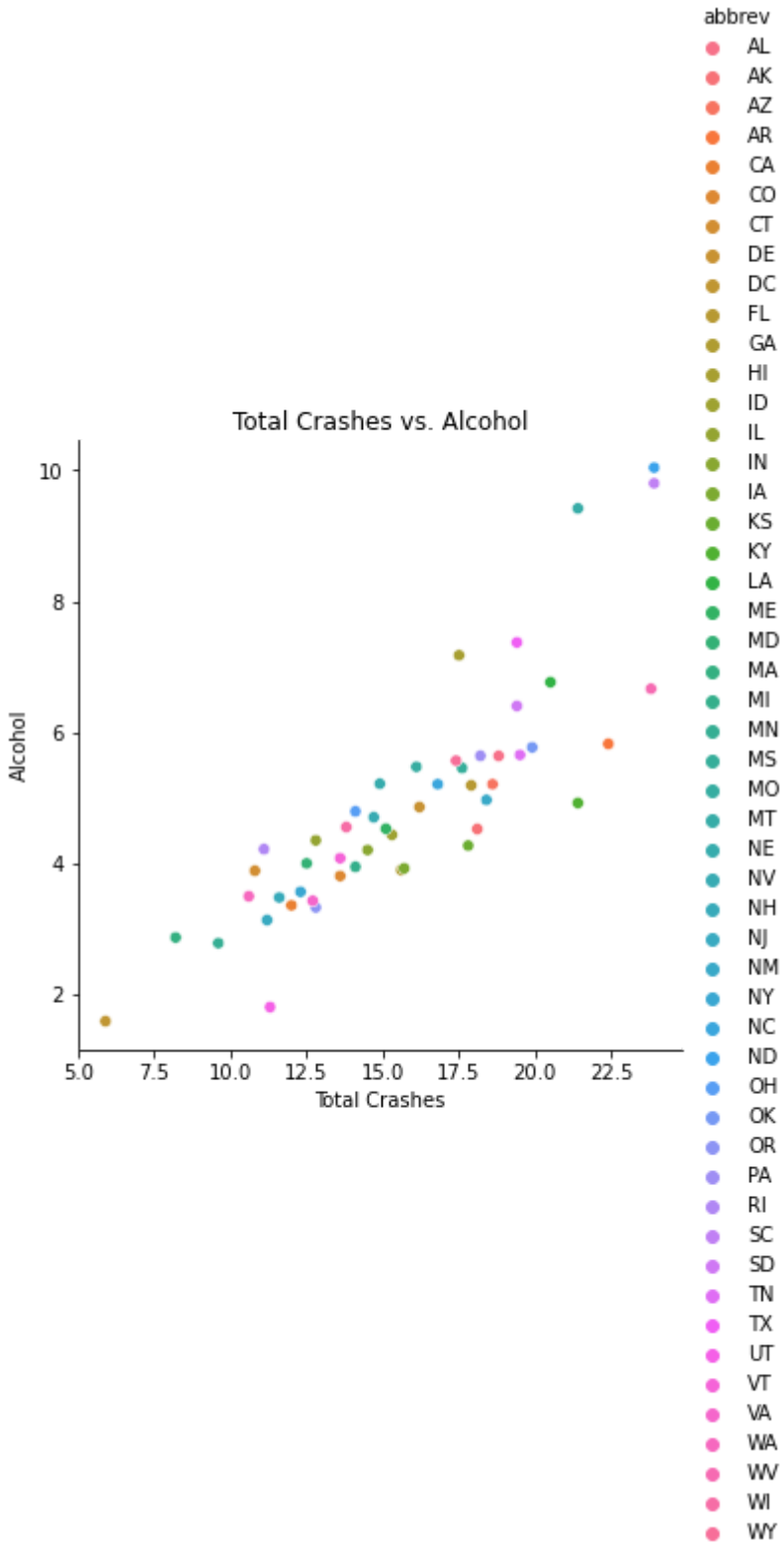

Total Crashes vs. Speeding
Speeding
Total Crashes
2
3
4
5
6
7
8
9
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
abbrev
AL
AK
AZ
AR
CA
CO
CT
DE
DC
FL
GA
HI
ID
IL
IN
IA
KS
KY
LA
ME
MD
MA
MI
MN
MS
MO
MT
NE
NV
NH
NJ
NM
NY
NC
ND
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VT
VA
WA
WV
WI
WY

In [32]:

```
sns.relplot(x="total",y="alcohol",data=dataset,hue="abbrev")
plt.ylabel('Alcohol')
plt.xlabel('Total Crashes')
plt.title('Total Crashes vs. Alcohol')

"""
Inference :
The relational plot ("relplot") illustrates the relationship between total car crashes a
Each point on the plot represents a data point in the dataset, and different states are
The plot provides a visual comparison of how alcohol-related crashes vary with the total
There isn't a clear linear trend in the relationship; points are scattered without a dis
"""
```

Out[32]:

```
'\nInference :\nThe relational plot ("relplot") illustrates the relations
hip between total car crashes and crashes involving alcohol.\nEach point
on the plot represents a data point in the dataset, and different states
are color-coded for comparison (hue).\nThe plot provides a visual compari
son of how alcohol-related crashes vary with the total number of crashes
in different states.\nThere isn\'t a clear linear trend in the relationsh
ip; points are scattered without a distinct pattern, suggesting that the
association between total crashes and alcohol-related incidents may diffe
r by state. Further state-specific analysis may be needed to explore this
further.\n'
```


Total Crashes vs. Alcohol
Alcohol
Total Crashes
10
8
6
4
2
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
abbrev
AL
AK
AZ
AR
CA
CO
CT
DE
DC
FL
GA
HI
ID
IL
IN
IA
KS
KY
LA
ME
MD
MA
MI
MN
MS
MO
MT
NE
NV
NH
NJ
NM
NY
NC
ND
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VT
VA
WA
WV
WI
WY

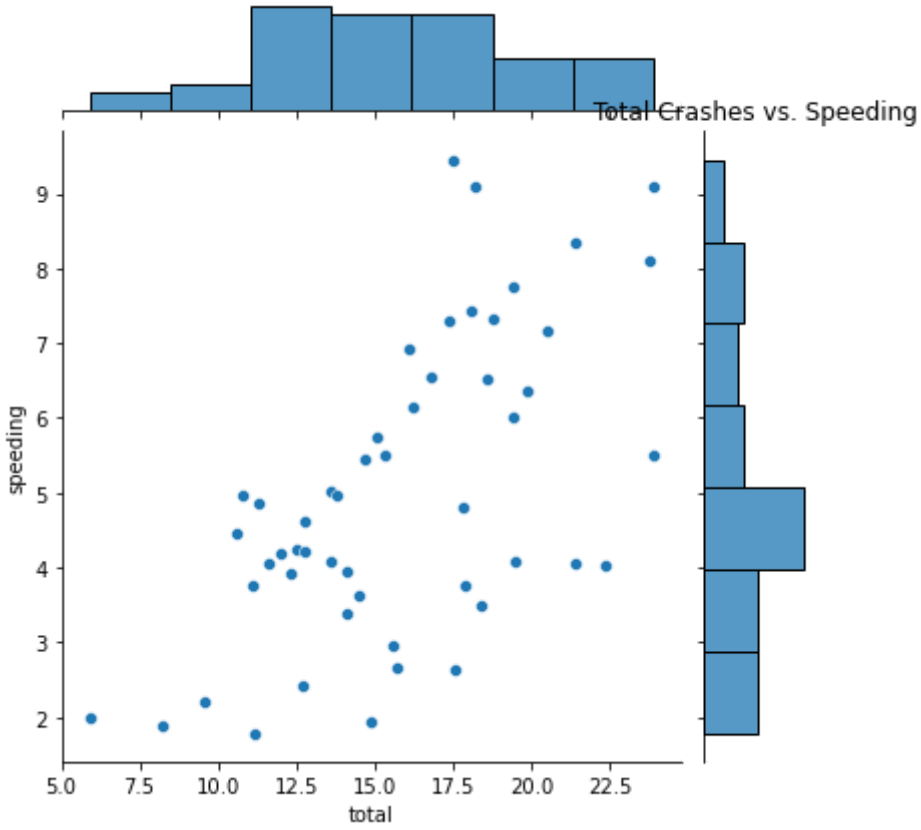
Jointplot

In [48]:

```
sns.jointplot(x="total",y="speeding",data=dataset)
plt.ylabel('Speeding')
plt.xlabel('Total Crashes')
plt.title('Total Crashes vs. Speeding')

"""
Inference :
The association between overall auto accidents and accidents involving speeding is shown
To show the distribution and correlation of the two variables, it combines a scatter plo
The scatter plot demonstrates that there isn't a clear linear correlation between overal
Additional details on the distributions of both variables are available from the histogr

"""
```

Out[48]:

```
"\nInference :\nThe association between overall auto accidents and accide
nts involving speeding is shown in the combined plot.\nTo show the distri
bution and correlation of the two variables, it combines a scatter plot w
ith histograms.\nThe scatter plot demonstrates that there isn't a clear l
inear correlation between overall crash rates and instances of speedin
g.\nAdditional details on the distributions of both variables are availab
le from the histograms on the top and right sides.\n\n"
```

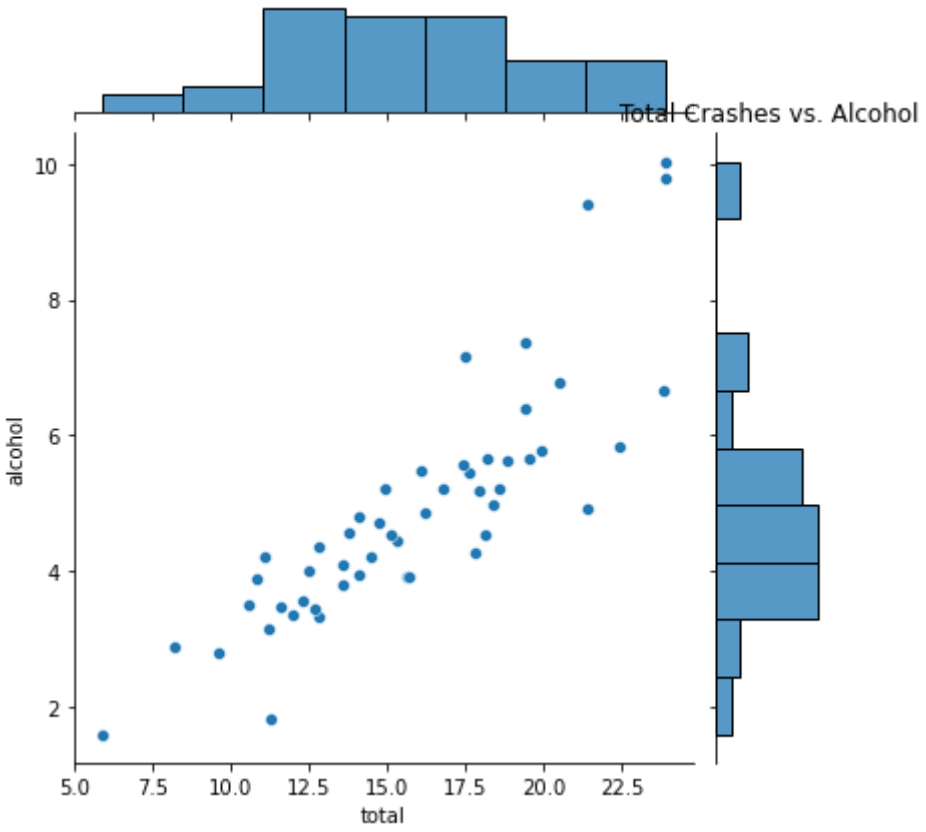


In [49]:

```
sns.jointplot(x="total",y="alcohol",data=dataset)
plt.ylabel('Alcohol')
plt.xlabel('Total Crashes')
plt.title('Total Crashes vs. Alcohol')

"""
Inference :
The association between overall auto accidents and accidents involving alcohol is repres
In order to shed light on the distribution and correlation between the two variables, it
The scatter plot demonstrates that there isn't a significant linear association between
Additional details on the distributions of both variables are provided by the histograms

"""
```

Out[49]:

```
"\nInference :\nThe association between overall auto accidents and accide
nts involving alcohol is represented by the joint plot.\nIn order to shed
light on the distribution and correlation between the two variables, it c
ombines a scatter plot and histograms.\nThe scatter plot demonstrates tha
t there isn't a significant linear association between overall crashes an
d instances involving alcohol.\nAdditional details on the distributions o
f both variables are provided by the histograms on the top and right side
s.\n\n"
```



Multivariate

Definition and Objective:

Multivariate refers to the analysis or study of three or more variables simultaneously in statistics. In multivariate analysis, researchers examine the relationships, interactions, patterns, and dependencies among multiple variables to gain a more comprehensive understanding of complex phenomena. This type of analysis is particularly useful when dealing with situations where several variables may be interconnected or when trying to

In [36]:

```
#Finding correlation between all attributes:
corr=dataset.corr()
corr
```

Out[36]:

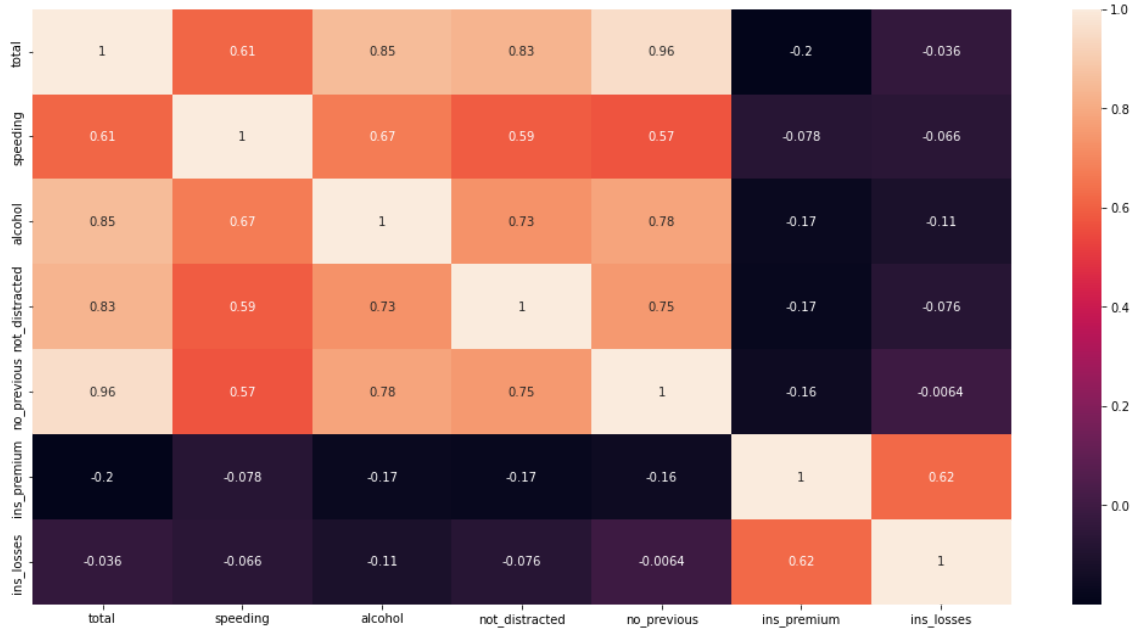
	total	speeding	alcohol	not_distracted	no_previous	ins_premium	in
total	1.000000	0.611548	0.852613	0.827560	0.956179	-0.199702	
speeding	0.611548	1.000000	0.669719	0.588010	0.571976	-0.077675	
alcohol	0.852613	0.669719	1.000000	0.732816	0.783520	-0.170612	
not_distracted	0.827560	0.588010	0.732816	1.000000	0.747307	-0.174856	
no_previous	0.956179	0.571976	0.783520	0.747307	1.000000	-0.156895	
ins_premium	-0.199702	-0.077675	-0.170612	-0.174856	-0.156895	1.000000	
ins_losses	-0.036011	-0.065928	-0.112547	-0.075970	-0.006359	0.623116	

In [50]:

```
plt.subplots(figsize=(18,9))
sns.heatmap(corr,annot=True)

"""
Inference :
The correlation between several dataset variables is represented visually by the heatmap
Lighter colours denote weaker or negative connections, while darker colours denote stron
Which variables are significantly connected and which are not can be quickly determined
For instance, a black cell denotes a significant positive correlation between two variab
The dataset's potential linkages and dependencies can be found using this visualisation.

"""
```

Out[50]:

```
"\nInference :\nThe correlation between several dataset variables is repr
esented visually by the heatmap.\nLighter colours denote weaker or negati
ve connections, while darker colours denote stronger positive correlation
s.\nWhich variables are significantly connected and which are not can be
quickly determined using the heatmap.\nFor instance, a black cell denotes
a significant positive correlation between two variables.\nThe dataset's
potential linkages and dependencies can be found using this visualisatio
n.\n\n"
```



In [ ]: