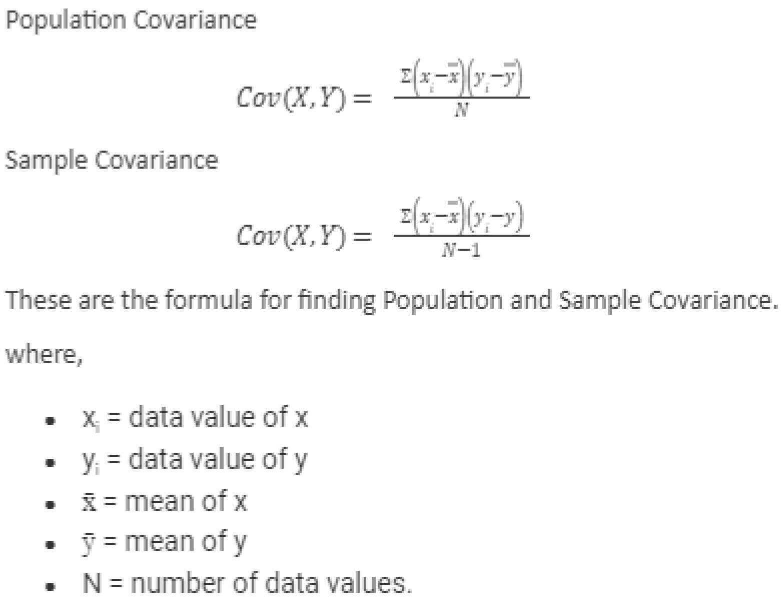
Understanding Covariance and Correlation-Elucidation and Illustration.

Covariance and correlation are both measures of the relationship between two variables, but they differ in terms of their scale and interpretation

Covariance:

Covariance is a measure of how much two random variables change together. It indicates the direction of the linear relationship between two variables but does not provide information about the strength of the relationship. Mathematically covariance can be computed using the below formula,



Population Covariance

$$Cov(X,Y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

$$Cov(X,Y) = \frac{\Sigma(x_i - \bar{x})(y_i - y)}{N-1}$$

These are the formula for finding Population and Sample Covariance.

where,

- x_i = data value of x
- y_i = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

- The value of covariance lies among -∞ and +∞
- A positive covariance indicates a positive relationship, where both variables tend to increase or decrease together. A negative covariance indicates an inverse relationship. If the covariance is zero, it indicates no linear relationship between the variables.
- Covariance does not provide a standardized measure of the strength of the relationship. Its value depends on the scale of the variables, making it difficult to compare covariances between different pairs of variables directly.

In [1]:
```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

In [10]:
```
# Load the tips dataset from Seaborn
tips = sns.load_dataset('tips')
print(tips.head(4))
```

```
   total_bill   tip     sex smoker  day    time  size
0       16.99  1.01  Female     No  Sun  Dinner     2
1       10.34  1.66    Male     No  Sun  Dinner     3
```

```
2        21.01  3.50     Male      No  Sun  Dinner     3
3        23.68  3.31     Male      No  Sun  Dinner     2
```

In [3]:
```
covariance = tips['total_bill'].cov(tips['tip'])
print("Covariance between 'total_bill' and 'tip':", covariance)
```

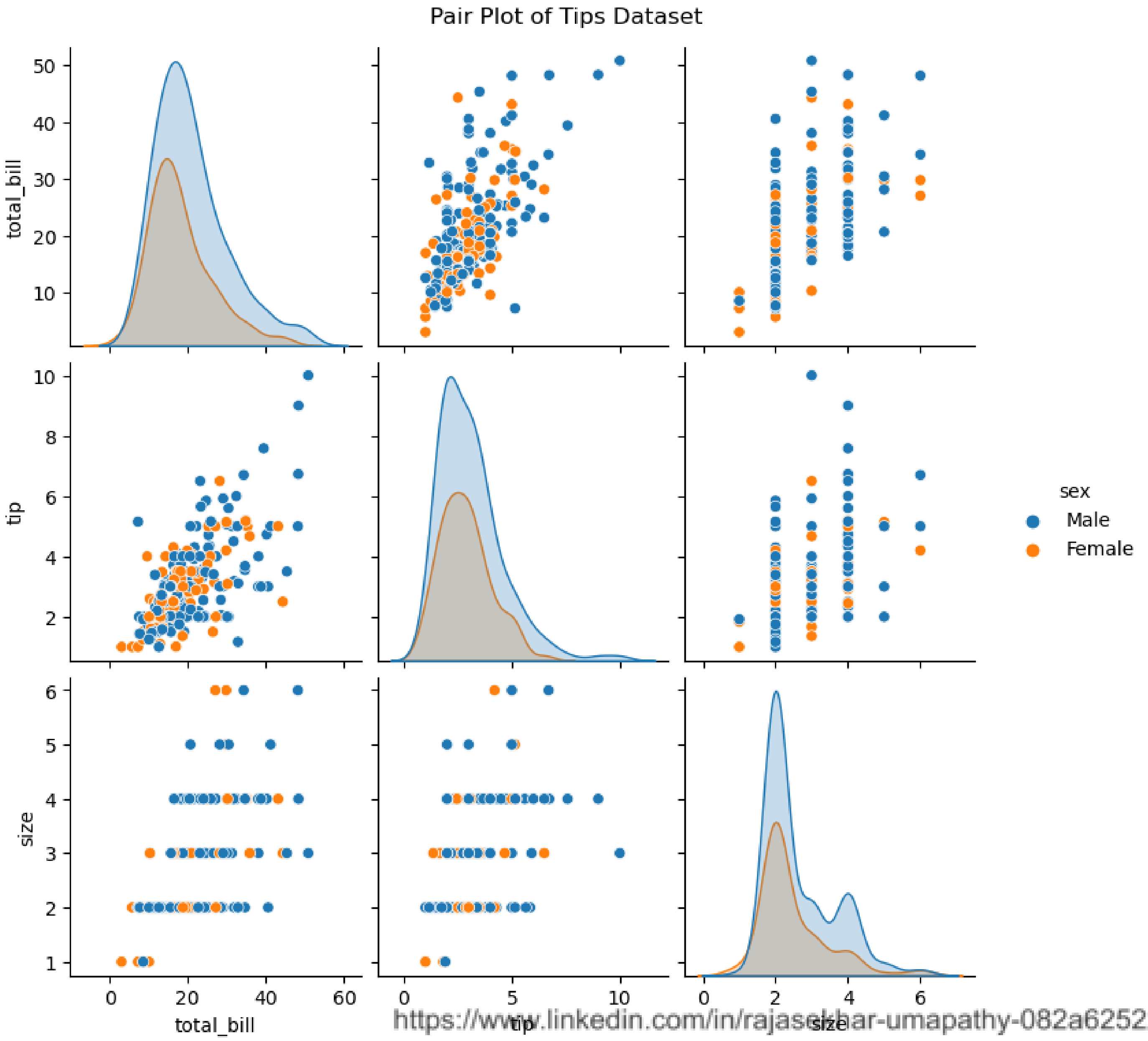
```
Covariance between 'total_bill' and 'tip': 8.323501629224854
```

The computed covariance value between the 'total_bill' and 'tip' columns represents the degree to which the total bill amount and tip amount change together. As the covariance value is positive, it suggests that as the total bill amount increases, the tip amount also tends to increase, and vice versa.

Let's try to visualize

Covariance is typically not visualized directly using plots because it is a single numerical value that represents the degree to which two variables change together. However, scatter plots can provide some insight into the covariance between variables. I will be using Pair plot in this example.

In [12]:
```
# Create a pair plot- It plots Scatter plots between numerical columns and histogram for
sns.pairplot(tips, hue='sex')
plt.suptitle("Pair Plot of Tips Dataset",y=1.02, fontsize=12, ha='center')
plt.show()
```

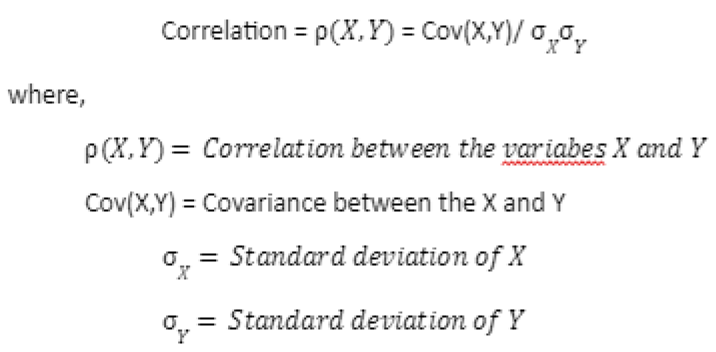


Inferences:

- We observe a positive correlation between total bill and tip amount, indicating that as the total bill increases, the tip amount tends to increase as well.
- There seems to be no strong relationship between total bill and party size.
- The pair plot also reveals differences in spending patterns between males and females.

Correlations:

Correlation is a standardized measure of the linear relationship between two variables. It quantifies both the direction and the strength of the relationship. Mathematically correlation can be computed using the below formula,



Correlation = $\rho(X, Y)$ = Cov(X,Y)/ $\sigma_X \sigma_Y$

where,

$\rho(X, Y) = $ *Correlation between the variabes X and Y*

Cov(X,Y) = Covariance between the X and Y

$\sigma_X = $ *Standard deviation of X*

$\sigma_Y = $ *Standard deviation of Y*

- Correlation coefficients range from -1 to 1
- Correlation coefficients have a clear interpretation. A correlation coefficient close to 1 or -1 indicates a strong linear relationship, while a coefficient close to 0 indicates a weak or no linear relationship.
- Correlation coefficients are unitless and standardized, making them comparable across different pairs of variables

In [6]:

```
# Calculate correlation between 'total_bill' and 'tip'
correlation = tips['total_bill'].corr(tips['tip'])
print("Correlation between 'total_bill' and 'tip':", correlation)
```

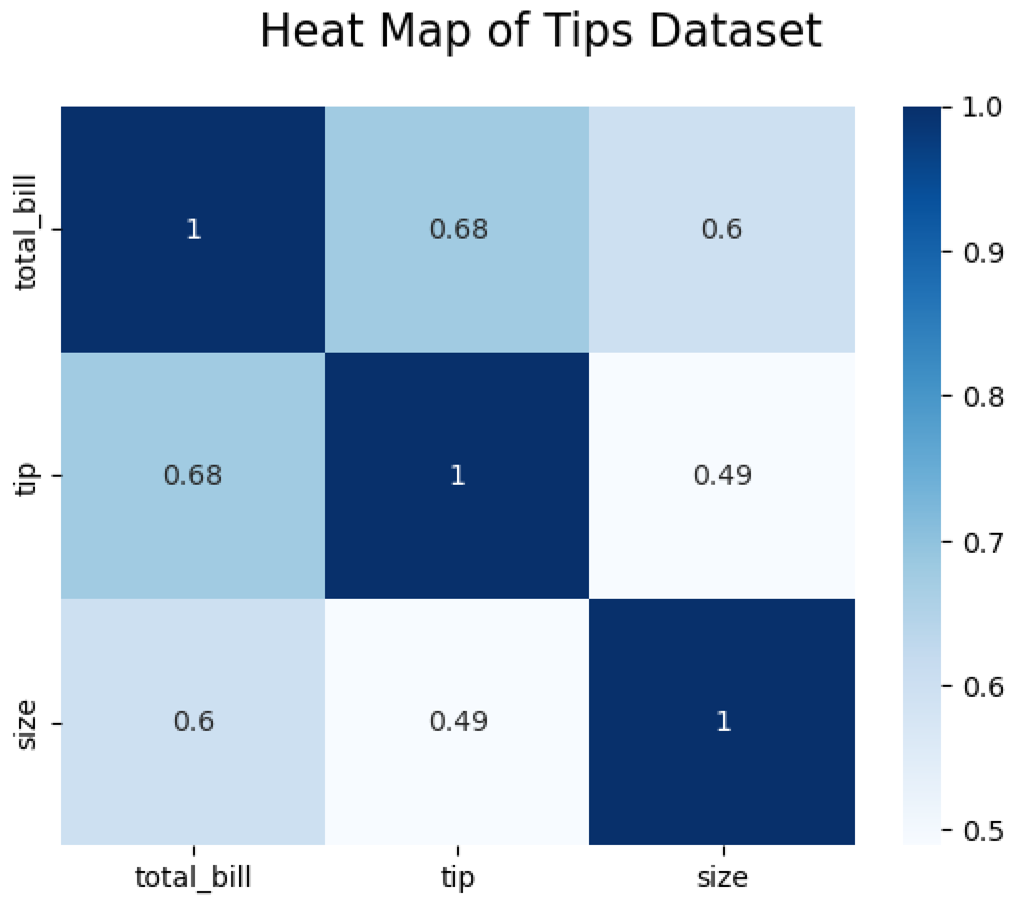
```
Correlation between 'total_bill' and 'tip': 0.6757341092113641
```

We observe a strong positive correlation between total bill and tip amount, with a correlation coefficient close to 0.68. (Stong positive as it is greater than +0.5).

Let's Visualize

In [7]:

```
correlation_matrix = tips.corr()
plt.suptitle("Heat Map of Tips Dataset", fontsize=16, ha='center')
sns.heatmap(correlation_matrix, annot=True, cmap='Blues')
plt.show()
```



Inferences:

- Strong positive correlation between total bill and tip amount, with a correlation coefficient as 0.68.
- Strong positive correlation between total bill and party size, with a correlation coefficient as 0.6.
- Weak positive correlation between party size and tip amount, with a correlation coefficient as 0.49

In summary, while covariance and correlation both measure the relationship between two variables, correlation provides a more interpretable and standardized measure of the strength and direction of the linear relationship. Correlation is preferred over covariance when comparing the relationships between variables, as it is less affected by differences in the scale of the variables.

https://www.linkedin.com/in/rajasekhar-umapathy-082a6252/