In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Data Set

https://www.kaggle.com/datasets/kolawale/focusing-on-mobile-app-or-website (https://www.kaggle.com/datasets/kolawale/focusing-on-mobile-app-or-website)

In [2]:
```
df = pd.read_csv('Ecus.csv') # import dataset
```

In [3]:
```
df
```

Out[3]:

	Email	Address	Avatar	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
0	mstephenson@fernandez.com	835 Frank Tunnel\nWrightmouth, MI 82180-9605	Violet	34.497268	12.655651	39.577668	4.082621	587.951054
1	hduke@hotmail.com	4547 Archer Common\nDiazchester, CA 06566-8576	DarkGreen	31.926272	11.109461	37.268959	2.664034	392.204933
2	pallen@yahoo.com	24645 Valerie Unions Suite 582\nCobbborough, D...	Bisque	33.000915	11.330278	37.110597	4.104543	487.547505
3	riverarebecca@gmail.com	1414 David Throughway\nPort Jason, OH 22070-1220	SaddleBrown	34.305557	13.717514	36.721283	3.120179	581.852344
4	mstephens@davidson-herman.com	14023 Rodriguez Passage\nPort Jacobville, PR 3...	MediumAquaMarine	33.330673	12.795189	37.536653	4.446308	599.406092
...	...	...	...	...	...	...	...	...
495	lewisjessica@craig-evans.com	4483 Jones Motorway Suite 872\nLake Jamiefurt,...	Tan	33.237660	13.566160	36.417985	3.746573	573.847438
496	katrina56@gmail.com	172 Owen Divide Suite 497\nWest Richard, CA 19320	PaleVioletRed	34.702529	11.695736	37.190268	3.576526	529.049004
497	dale88@hotmail.com	0787 Andrews Ranch Apt. 633\nSouth Chadburgh, ...	Cornsilk	32.646777	11.499409	38.332576	4.958264	551.620145
498	cwilson@hotmail.com	680 Jennifer Lodge Apt. 808\nBrendachester, TX...	Teal	33.322501	12.391423	36.840086	2.336485	456.469510
499	hannahwilson@davidson.com	49791 Rachel Heights Apt. 898\nEast Drewboroug...	DarkMagenta	33.715981	12.418808	35.771016	2.735160	497.778642

500 rows × 8 columns

In [4]:
```
df.head(10)
```

Out[4]:

	Email	Address	Avatar	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
0	mstephenson@fernandez.com	835 Frank Tunnel\nWrightmouth, MI 82180-9605	Violet	34.497268	12.655651	39.577668	4.082621	587.951054
1	hduke@hotmail.com	4547 Archer Common\nDiazchester, CA 06566-8576	DarkGreen	31.926272	11.109461	37.268959	2.664034	392.204933
2	pallen@yahoo.com	24645 Valerie Unions Suite 582\nCobbborough, D...	Bisque	33.000915	11.330278	37.110597	4.104543	487.547505
3	riverarebecca@gmail.com	1414 David Throughway\nPort Jason, OH 22070-1220	SaddleBrown	34.305557	13.717514	36.721283	3.120179	581.852344
4	mstephens@davidson-herman.com	14023 Rodriguez Passage\nPort Jacobville, PR 3...	MediumAquaMarine	33.330673	12.795189	37.536653	4.446308	599.406092
5	alvareznancy@lucas.biz	645 Martha Park Apt. 611\nJeffreychester, MN 6...	FloralWhite	33.871038	12.026925	34.476878	5.493507	637.102448
6	katherine20@yahoo.com	68388 Reyes Lights Suite 692\nJosephbury, WV 9...	DarkSlateBlue	32.021596	11.366348	36.683776	4.685017	521.572175
7	awatkins@yahoo.com	Unit 6538 Box 8980\nDPO AP 09026-4941	Aqua	32.739143	12.351959	37.373359	4.434273	549.904146
8	vchurch@walter-martinez.com	860 Lee Key\nWest Debra, SD 97450-0495	Salmon	33.987773	13.386235	37.534497	3.273434	570.200409
9	bonnie69@lin.biz	PSC 2734, Box 5255\nAPO AA 98456-7482	Brown	31.936549	11.814128	37.145168	3.202806	427.199385

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Email                 500 non-null    object 
 1   Address               500 non-null    object 
 2   Avatar                500 non-null    object 
 3   Avg. Session Length   500 non-null    float64
 4   Time on App           500 non-null    float64
 5   Time on Website       500 non-null    float64
 6   Length of Membership  500 non-null    float64
 7   Yearly Amount Spent   500 non-null    float64
dtypes: float64(5), object(3)
memory usage: 31.4+ KB
```

In [6]:
```
df.describe()
```

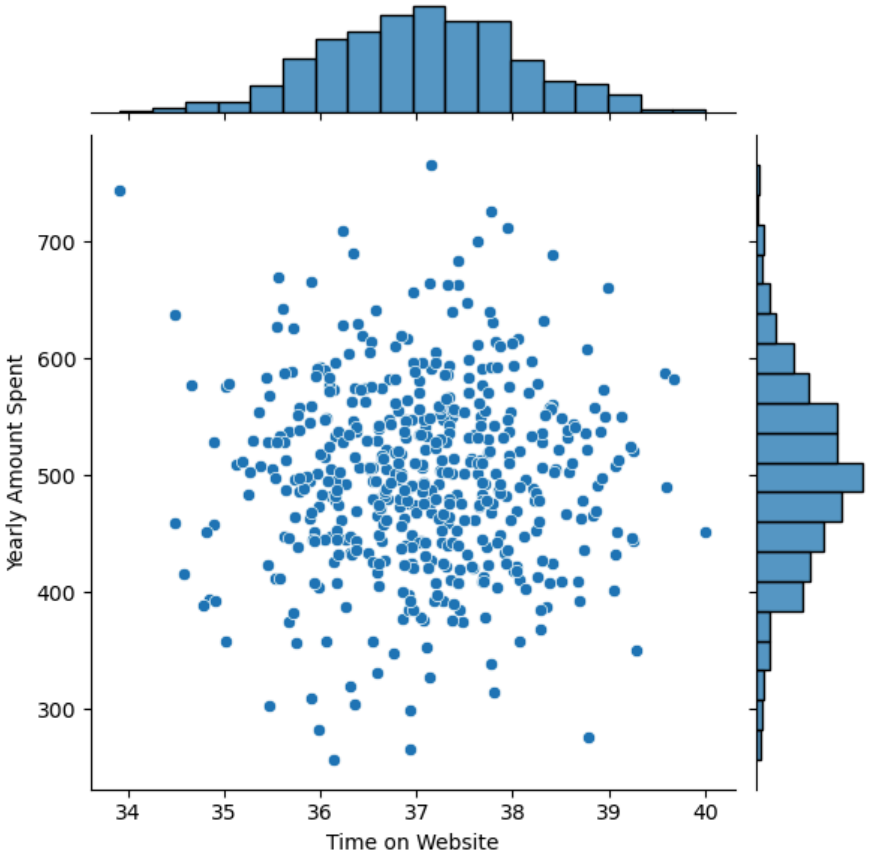
Out[6]:

	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
count	500.000000	500.000000	500.000000	500.000000	500.000000
mean	33.053194	12.052488	37.060445	3.533462	499.314038
std	0.992563	0.994216	1.010489	0.999278	79.314782
min	29.532429	8.508152	33.913847	0.269901	256.670582
25%	32.341822	11.388153	36.349257	2.930450	445.038277
50%	33.082008	11.983231	37.069367	3.533975	498.887875
75%	33.711985	12.753850	37.716432	4.126502	549.313828
max	36.139662	15.126994	40.005182	6.922689	765.518462

EDA

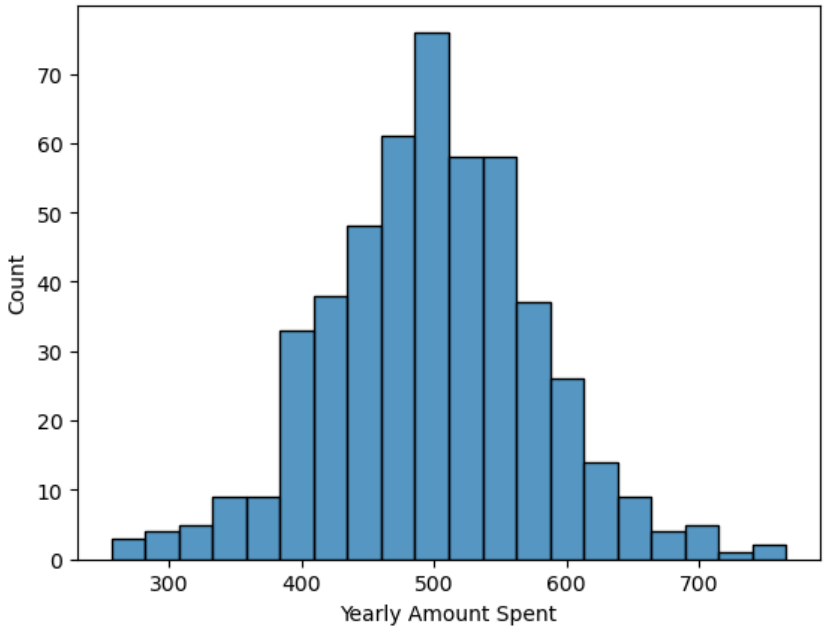
In [15]:
```
sns.jointplot(x= 'Time on Website', y= 'Yearly Amount Spent', data=df) #To check if there are any Correlation
```

Out[15]: <seaborn.axisgrid.JointGrid at 0x2628d2d0150>



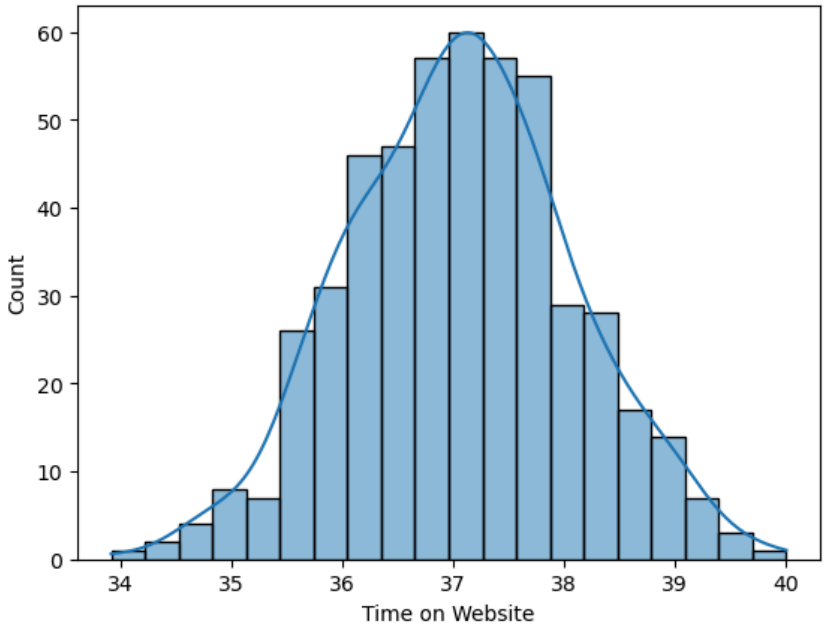
In [16]:
```
sns.histplot(df['Yearly Amount Spent'],bins = 20,kde= False)
```

Out[16]: <Axes: xlabel='Yearly Amount Spent', ylabel='Count'>



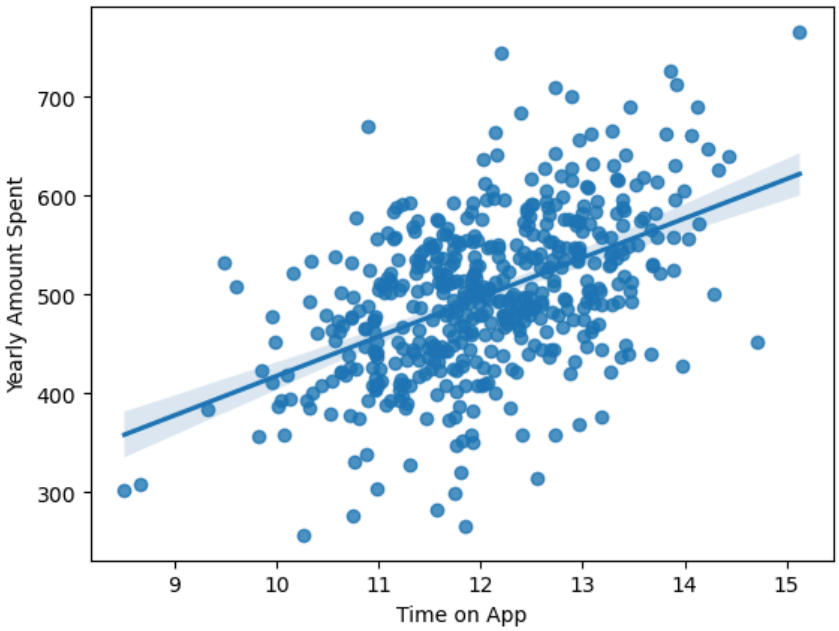
In [17]:
```
sns.histplot(df['Time on Website'],bins = 20,kde= True)
```

Out[17]:
```
<Axes: xlabel='Time on Website', ylabel='Count'>
```



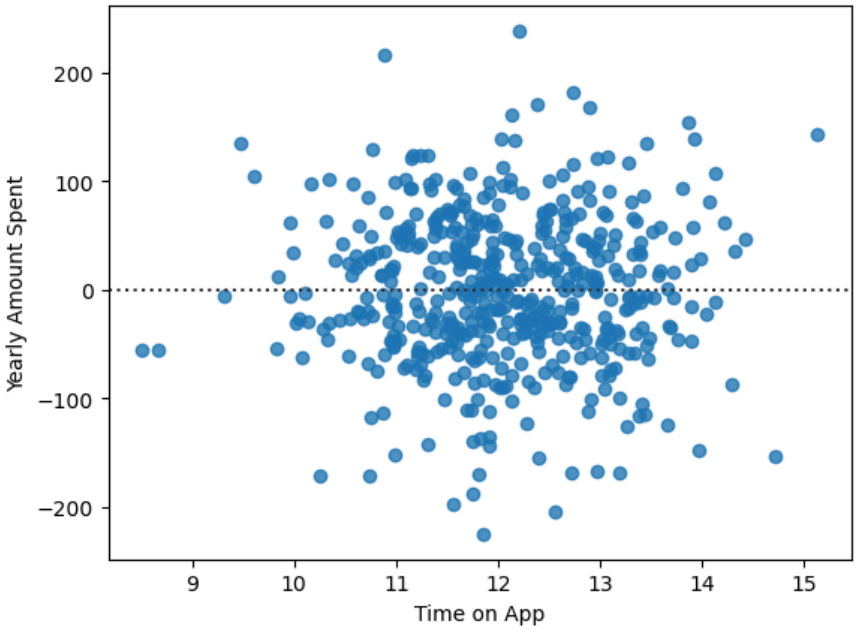
In [18]:
```
sns.regplot(x='Time on App', y= 'Yearly Amount Spent', data=df)
```

Out[18]:
```
<Axes: xlabel='Time on App', ylabel='Yearly Amount Spent'>
```



In [19]:
```
sns.residplot(x='Time on App', y= 'Yearly Amount Spent', data=df)
```

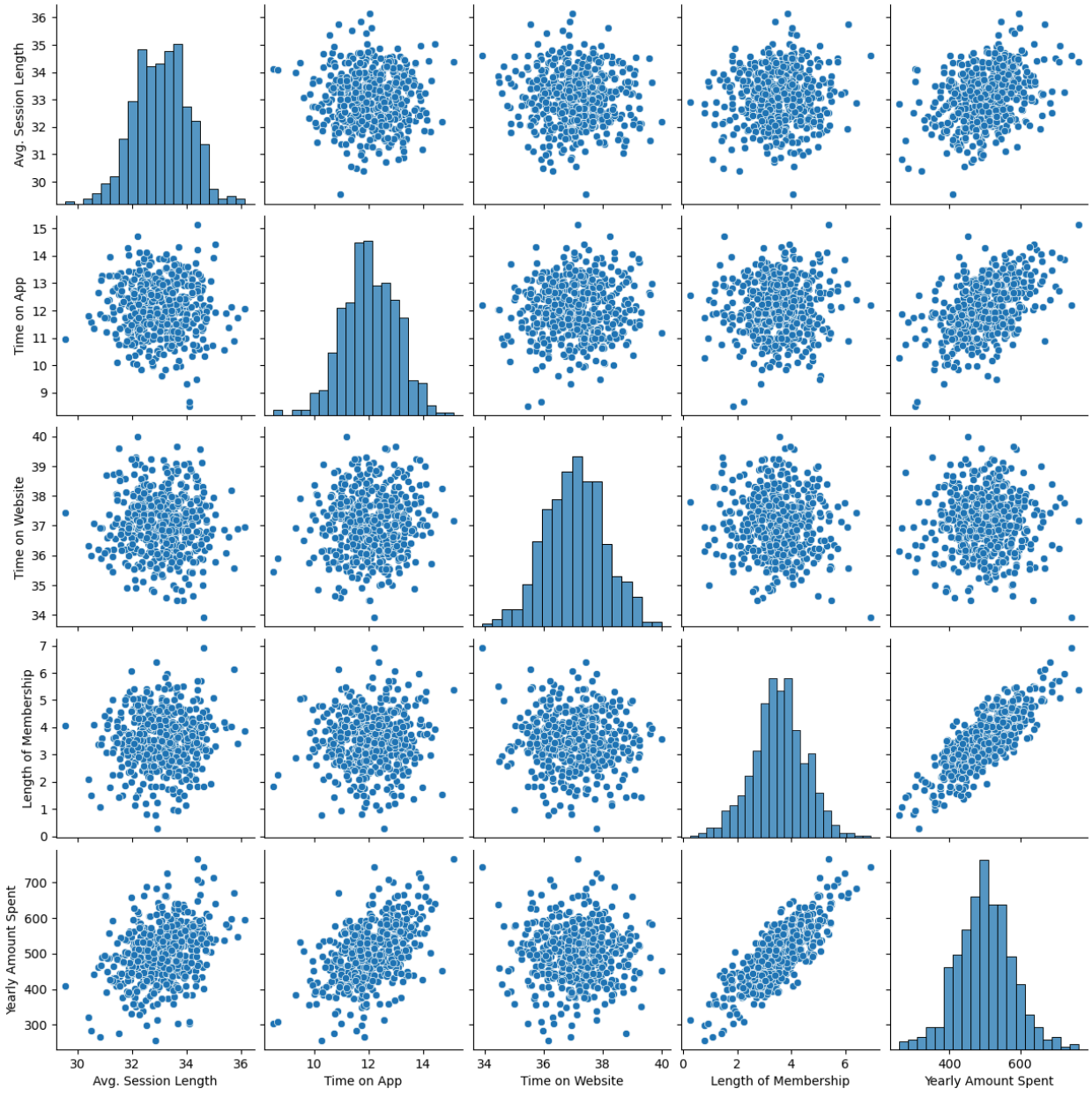
Out[19]:
```
<Axes: xlabel='Time on App', ylabel='Yearly Amount Spent'>
```



In [26]:
```
sns.pairplot(df)
```

```
C:\Users\lahir\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The figure layout has chang
ed to tight
  self._figure.tight_layout(*args, **kwargs)
```

Out[26]: <seaborn.axisgrid.PairGrid at 0x26297c52d50>



In [21]:
```
df.head()
```

Out[21]:

	Email	Address	Avatar	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
0	mstephenson@fernandez.com	835 Frank Tunnel\nWrightmouth, MI 82180-9605	Violet	34.497268	12.655651	39.577668	4.082621	587.951054
1	hduke@hotmail.com	4547 Archer Common\nDiazchester, CA 06566-8576	DarkGreen	31.926272	11.109461	37.268959	2.664034	392.204933
2	pallen@yahoo.com	24645 Valerie Unions Suite 582\nCobbborough, D...	Bisque	33.000915	11.330278	37.110597	4.104543	487.547505
3	riverarebecca@gmail.com	1414 David Throughway\nPort Jason, OH 22070-1220	SaddleBrown	34.305557	13.717514	36.721283	3.120179	581.852344
4	mstephens@davidson-herman.com	14023 Rodriguez Passage\nPort Jacobville, PR 3...	MediumAquaMarine	33.330673	12.795189	37.536653	4.446308	599.406092

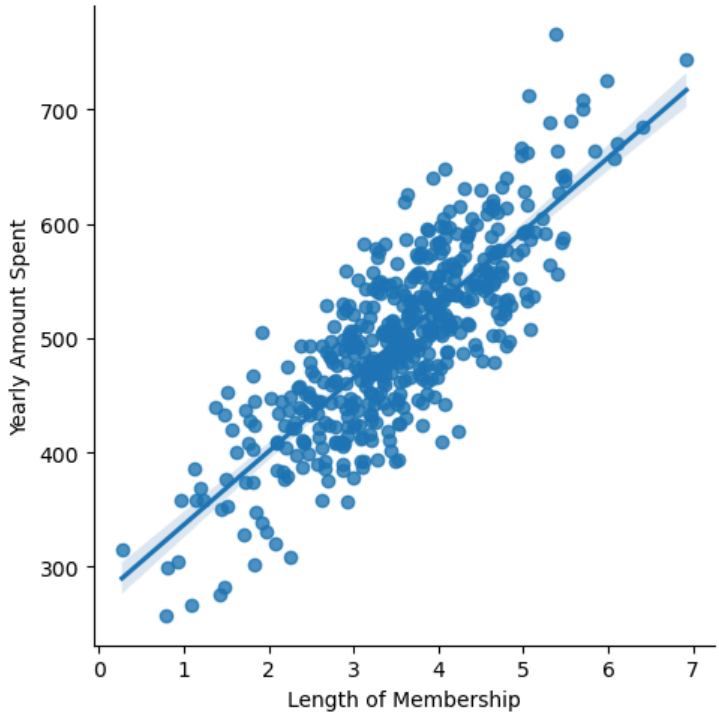
In [27]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   Email                 500 non-null    object 
 1   Address               500 non-null    object 
 2   Avatar                500 non-null    object 
 3   Avg. Session Length   500 non-null    float64
 4   Time on App           500 non-null    float64
 5   Time on Website       500 non-null    float64
 6   Length of Membership  500 non-null    float64
 7   Yearly Amount Spent   500 non-null    float64
dtypes: float64(5), object(3)
memory usage: 31.4+ KB
```

In [28]:
```
sns.lmplot(x = 'Length of Membership',
           y = 'Yearly Amount Spent',
           data = df)
```

```
C:\Users\lahir\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The figure layout has chang
ed to tight
  self._figure.tight_layout(*args, **kwargs)
```

Out[28]: <seaborn.axisgrid.FacetGrid at 0x2628db542d0>



In [29]:
```
from sklearn.model_selection import train_test_split
```

In [33]:
```
X = df[['Avg. Session Length','Time on App','Time on Website','Length of Membership']]
y = df['Yearly Amount Spent']
```

In [34]:
```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2,random_state=42)
```

Training The Model

In [41]:
```
from sklearn.linear_model import LinearRegression
```

In [42]:
```
lm = LinearRegression()
```

In [43]:
```
lm.fit(X_train,y_train)
```

Out[43]:
```
LinearRegression
LinearRegression()
```

In [44]:
```
lm.coef_
```

Out[44]:
```
array([25.5962591 , 38.78534598,  0.31038593, 61.89682859])
```

In [51]:
```
cdf = pd.DataFrame(lm.coef_,X.columns,columns = ['Coef'])
```

In [52]:
```
print(cdf)
```

```
                        Coef
Avg. Session Length   25.596259
Time on App           38.785346
Time on Website        0.310386
Length of Membership  61.896829
```

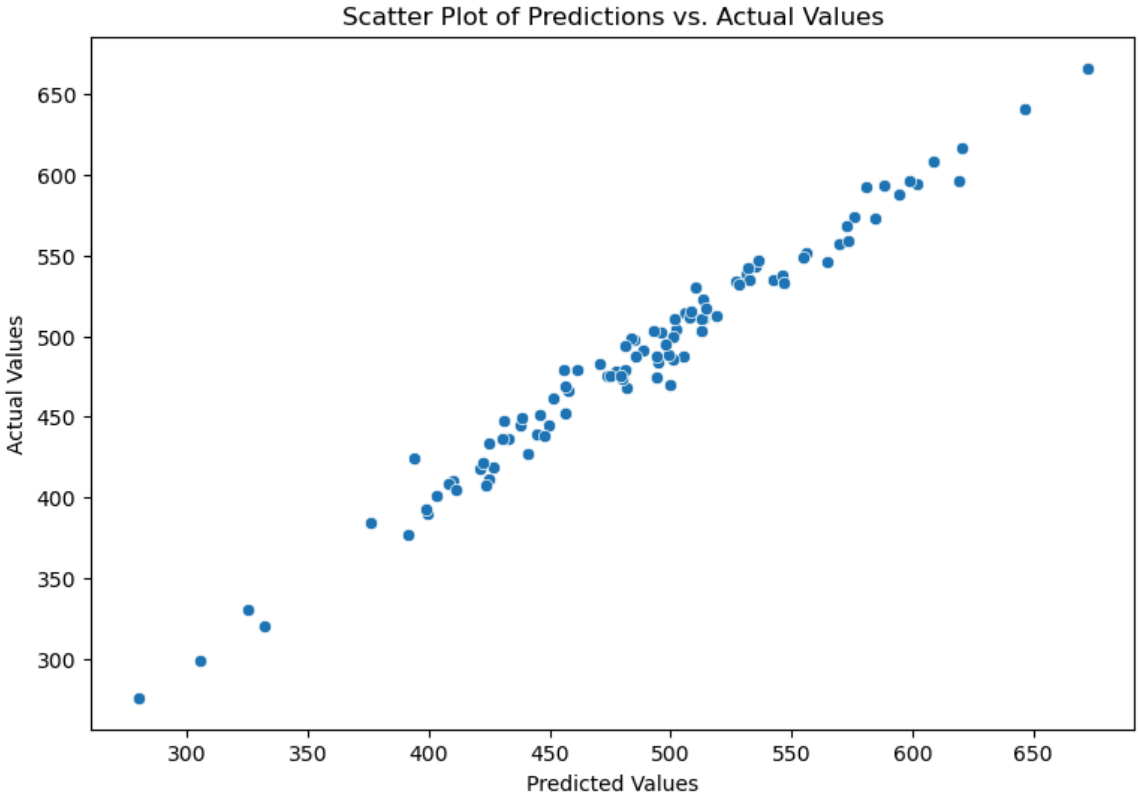
Predictions

In [53]:
```
prediction = lm.predict(X_test)
```

In [54]:
```
prediction
```

Out[54]:
```
array([402.86230051, 542.53325708, 426.62011918, 501.91386363,
       409.6666551 , 569.92155038, 531.50423529, 505.94309188,
       408.10378607, 473.45942928, 441.18668812, 424.52463471,
       424.83341694, 527.12061508, 430.87985533, 423.47062047,
       575.8751518 , 484.6563331 , 457.77896975, 481.58742311,
       501.56110993, 513.12815188, 507.49166899, 646.63377343,
       449.70050586, 496.26290484, 556.18523776, 554.78684161,
       399.1582784 , 325.16921284, 532.62732659, 477.73025415,
       500.76491535, 305.09971374, 505.46811902, 483.52069444,
       519.09464122, 437.75549737, 456.25005245, 470.63517876,
       494.11207805, 444.65549239, 508.57079732, 500.88197484,
       488.35128728, 535.34025218, 594.58301773, 513.59474408,
       279.69877702, 432.71590835, 421.06976164, 480.94327496,
       584.59481888, 608.61734059, 564.42312991, 494.47224504,
       393.95593318, 456.11321352, 572.92228417, 499.27385693,
       512.42973545, 391.56170305, 479.60705887, 481.05023229,
       474.71926117, 546.37716047, 430.11675694, 601.91418143,
       422.26508516, 493.11622454, 528.10614863, 581.06630842,
       620.60774498, 512.47838603, 411.2147464 , 498.07095351,
       461.44587681, 445.63453258, 447.63898998, 534.81030495,
       598.85091016, 619.46554961, 494.43362232, 672.2442837 ,
       532.15516513, 438.41740681, 514.80907179, 546.73893548,
       331.73069072, 510.33949236, 536.21660556, 499.50696031,
       375.86919792, 573.61952185, 479.18212334, 588.32862943,
       485.18137257, 455.93070091, 398.67820721, 451.70869105])
```

In [63]:
```
plt.figure(figsize=(9, 6))
sns.scatterplot(x=prediction, y=y_test)
plt.title('Scatter Plot of Predictions vs. Actual Values')
plt.xlabel('Predicted Values')
plt.ylabel('Actual Values')
plt.show()
```



Residual Analysis

In [66]:
```
from sklearn.metrics import mean_squared_error,mean_absolute_error
import math
```

In [70]:
```
print("Mean Square Error",mean_squared_error(y_test,prediction))
print("Mean Absolute Error",mean_absolute_error(y_test,prediction))
```

```
Mean Square Error 109.86374118393988
Mean Absolute Error 8.558441885315233
```

In [71]:
```
residuals = y_test - prediction
```

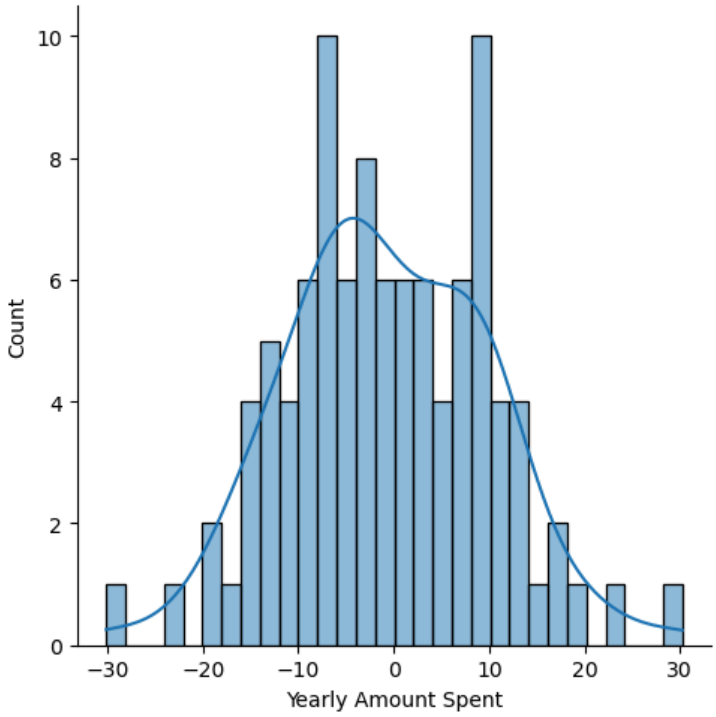
In [72]:
```
residuals
```

Out[72]:
```
361    -1.829165
73     -7.756069
374    -8.017377
155     2.064515
104     0.402956
          ...
347     4.827772
86      2.197933
75     22.788656
438    -5.685951
15     10.072051
Name: Yearly Amount Spent, Length: 100, dtype: float64
```

In [74]:
```
sns.displot(residuals,bins=30,kde=True)
```

```
C:\Users\lahir\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The figure layout has chang
ed to tight
  self._figure.tight_layout(*args, **kwargs)
```

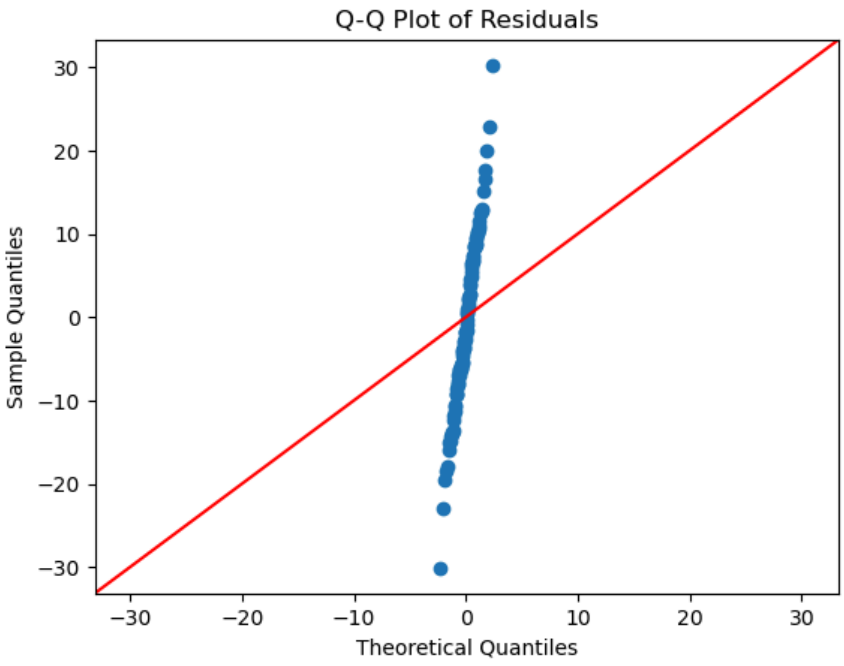
Out[74]: <seaborn.axisgrid.FacetGrid at 0x26299af6a50>



In [75]:
```
import statsmodels.api as sm
```

In [84]:
```
plt.figure(figsize=(10, 6))
sm.qqplot(residuals, line = '45')
plt.title('Q-Q Plot of Residuals')
plt.xlabel('Theoretical Quantiles')
plt.ylabel('Sample Quantiles')
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```



Thank you !

By Lahiru Sadakelum