Linear Regression Project

Congratulations! You just got some contract work with an Ecommerce company based in New York City that sells clothing online but they also have in-store style and clothing advice sessions. Customers come in to the store, have sessions/meetings with a personal stylist, then they can go home and order either on a mobile app or website for the clothes they want.

The company is trying to decide whether to focus their efforts on their mobile app experience or their website. They've hired you on contract to help them figure it out! Let's get started!

Just follow the steps below to analyze the customer data (it's fake, don't worry I didn't give you real credit card numbers or emails).

Imports

Import pandas, numpy, matplotlib,and seaborn. Then set %matplotlib inline (You'll import sklearn as you need it.)

In [2]:
```
#Lets import all our libraries

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

In [3]:
```
#Import matplotline

%matplotlib inline
```

Get the Data

We'll work with the Ecommerce Customers csv file from the company. It has Customer info, suchas Email, Address, and their color Avatar. Then it also has numerical value columns:

- Avg. Session Length: Average session of in-store style advice sessions.
- Time on App: Average time spent on App in minutes
- Time on Website: Average time spent on Website in minutes
- Length of Membership: How many years the customer has been a member.

Read in the Ecommerce Customers csv file as a DataFrame called customers.

In [32]:
```
customers = pd.read_csv('Ecommerce Customers')
```

Check the head of customers, and check out its info() and describe() methods.

In [33]:
```
customers.head()
```

Out[33]:

	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 [34]:
```
customers.describe()
```

Out[34]:

	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

In [35]:
```
customers.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
```

Exploratory Data Analysis

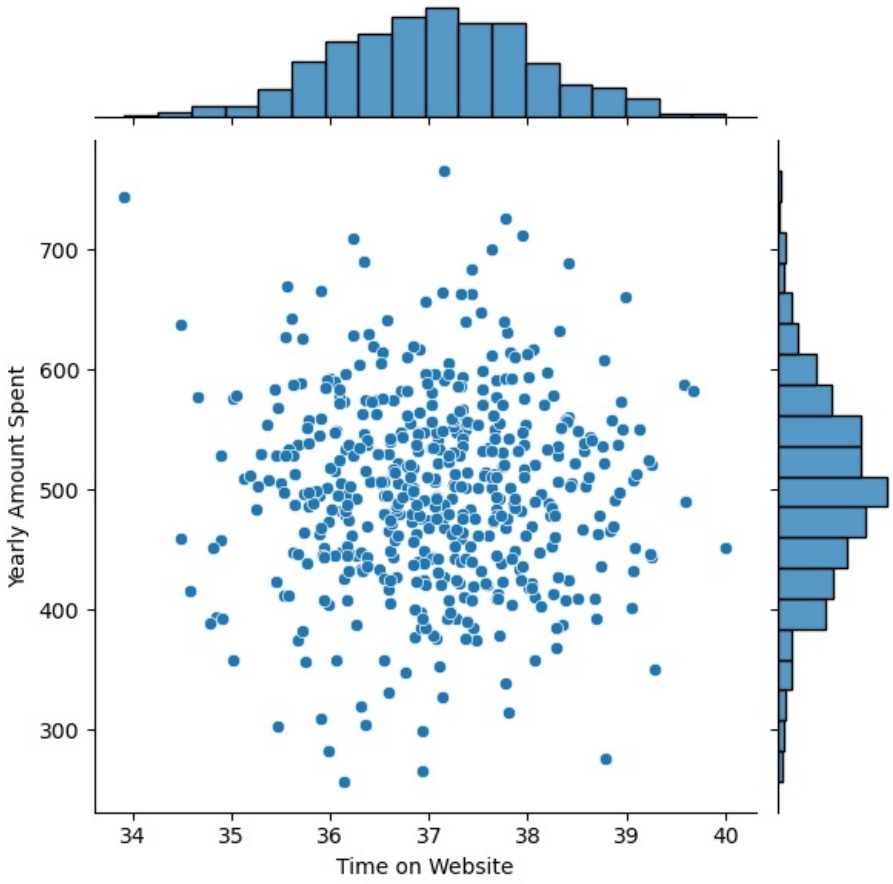
Let's explore the data!

For the rest of the exercise we'll only be using the numerical data of the csv file.

Use seaborn to create a jointplot to compare the Time on Website and Yearly Amount Spent columns. Does the correlation make sense?

In [36]:
```
sns.jointplot(x='Time on Website',y='Yearly Amount Spent', data=customers)
```

Out[36]:
```
<seaborn.axisgrid.JointGrid at 0x2886eea6c70>
```

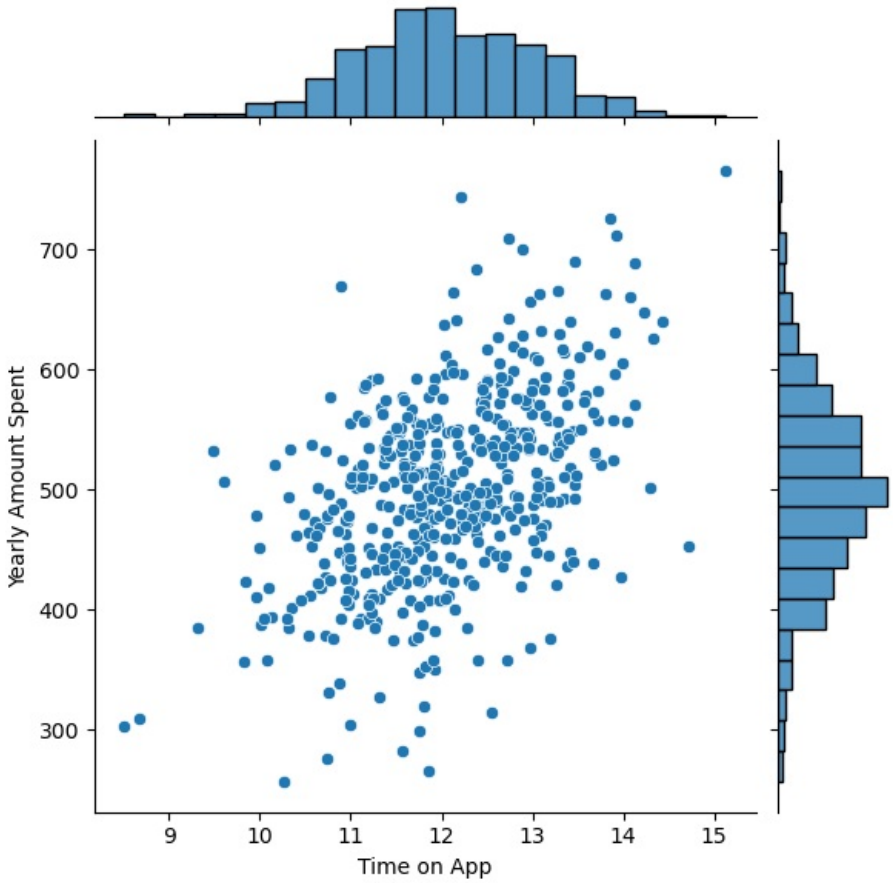


Do the same but with the Time on App column instead

Do the same but with the Time on App column instead.

In [37]:
```
sns.jointplot(x='Time on App',y='Yearly Amount Spent', data=customers)
```

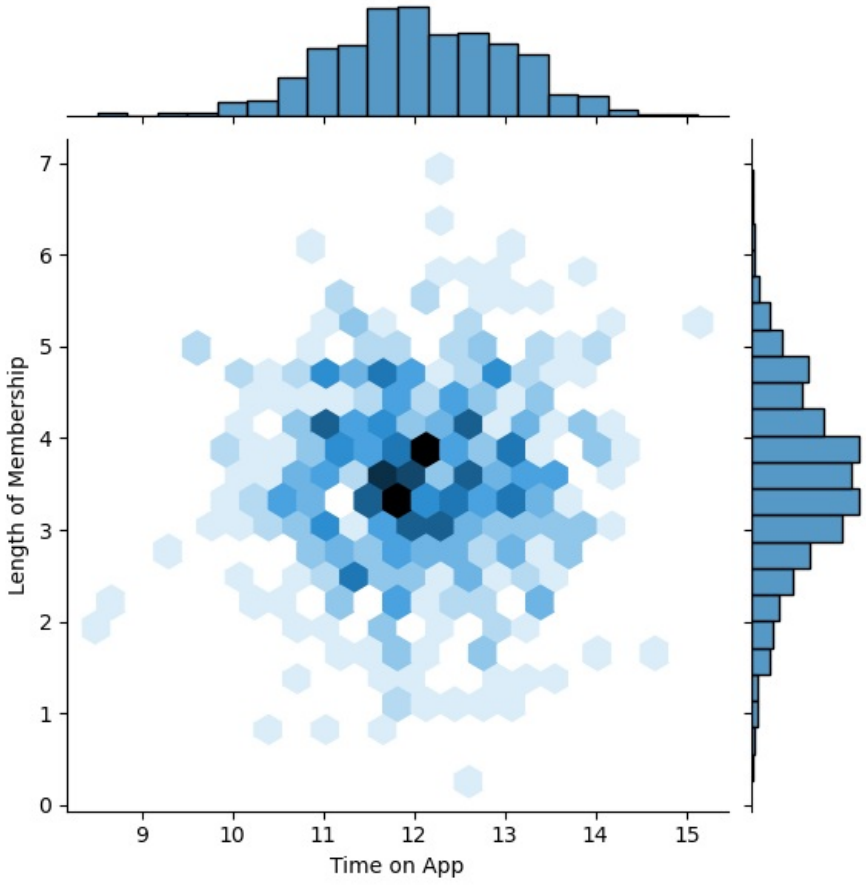
Out[37]:
```
<seaborn.axisgrid.JointGrid at 0x2886f23fbe0>
```



Use jointplot to create a 2D hex bin plot comparing Time on App and Length of Membership.

In [38]:
```
sns.jointplot(x='Time on App',y='Length of Membership', data=customers, kind='hex')
```

Out[38]:
```
<seaborn.axisgrid.JointGrid at 0x28872940e80>
```

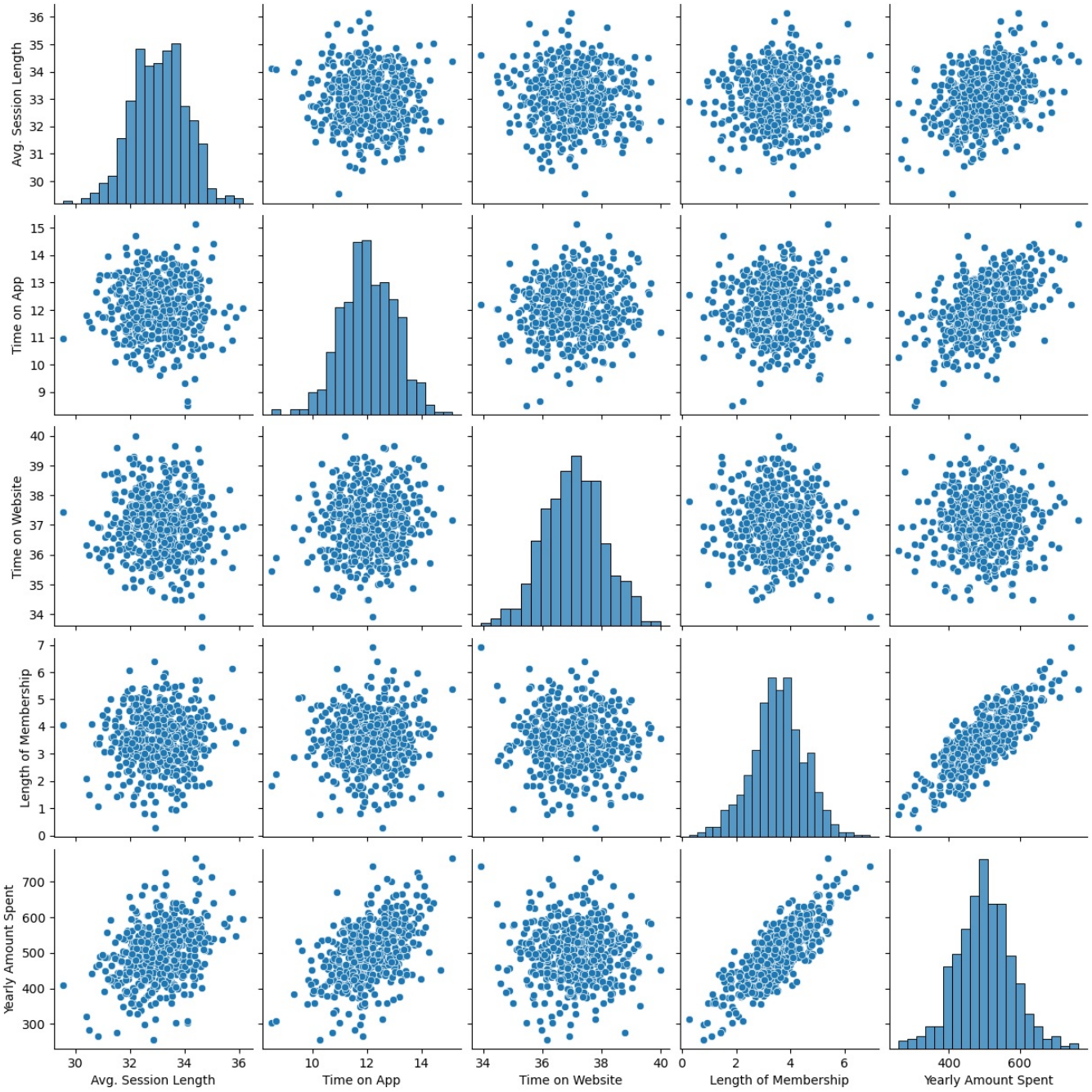


Let's explore these types of relationships across the entire data set. Use pairplot to recreate the plot below.(Don't worry about

the the colors)

In [41]:
```
sns.pairplot(customers)
```

Out[41]:
```
<seaborn.axisgrid.PairGrid at 0x28872534a90>
```



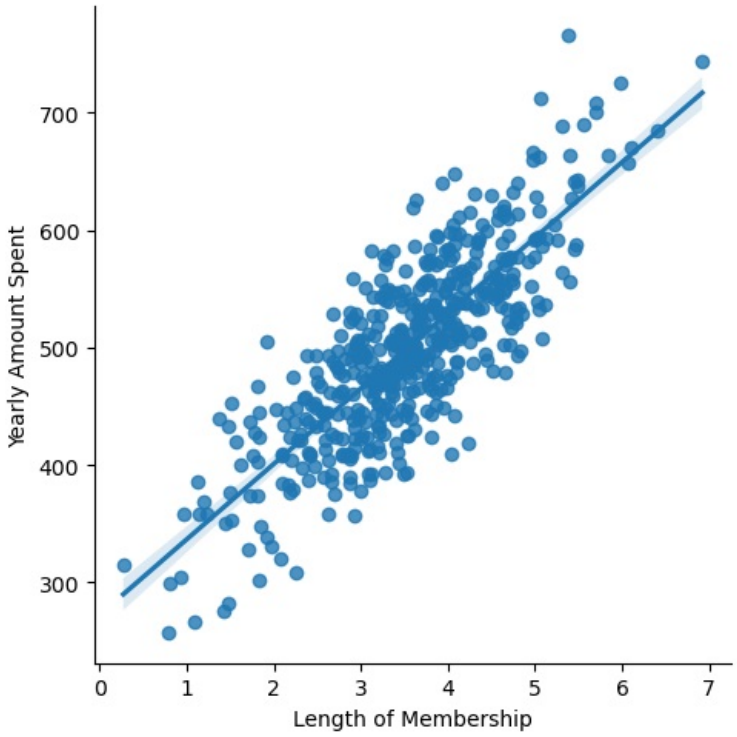
Based off this plot what looks to be the most correlated feature with Yearly Amount Spent?

In [285…
```
#The above analysis shows that it is the length of Membership.
# This is because the scatterplots somewhat form a diagonal line which is an indication of correlation.
```

Create a linear model plot (using seaborn's lmplot) of Yearly Amount Spent vs. Length of Membership.

In [42]:
```
sns.lmplot(x='Length of Membership',y='Yearly Amount Spent',data=customers)
```

Out[42]:
```
<seaborn.axisgrid.FacetGrid at 0x28874456e80>
```



Training and Testing Data

Now that we've explored the data a bit, let's go ahead and split the data into training and testing sets. **Set a variable X equal to the numerical features of the customers and a variable y equal to the "Yearly Amount Spent" column.**

In [50]:
```
#Let's get to building our X array with features to train on
# To get the columns to copy and paste to our X framework we do

customers.columns
```

Out[50]:
```
Index(['Email', 'Address', 'Avatar', 'Avg. Session Length', 'Time on App',
       'Time on Website', 'Length of Membership', 'Yearly Amount Spent'],
      dtype='object')
```

In [51]:
```
#We omitted Email, Address and Avatar as their dtype is an object and hence cannot be trained
# in the circumstances that are given to us.
#Let's setup our X_train array df

X = customers[['Avg. Session Length', 'Time on App',
       'Time on Website', 'Length of Membership']]
```

In [52]:
```
#Setting up y array which is our target variable

y = customers['Yearly Amount Spent']
```

In [53]:
```
#Now our next step is to develop a train test split on our model. We want to split
# our data in a training set and a testing set for our model once it has been trained.
```

Use model_selection.train_test_split from sklearn to split the data into training and testing sets. Set test_size=0.3 and random_state=101

In [54]:
```
#Let's import our train_test_split that will be used in our model building

from sklearn.model_selection import train_test_split
```

In [55]:
```
#Just to check the format
```

```
train_test_split
```

Out[55]:
```
<function sklearn.model_selection._split.train_test_split(*arrays, test_size=None, train_size=None, random_state=None, shuffle=True, stratify=None)>
```

In [56]:
```
#By pressing shift + tab, let's plug in the function below.
# Re-adjustment of the default test_size and random_state was done.
#The function is as follows;

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
```

Training the Model

Now its time to train our model on our training data!

Import LinearRegression from sklearn.linear_model

In [57]:
```
from sklearn.linear_model import LinearRegression
```

Create an instance of a LinearRegression() model named lm.

In [58]:
```
lm = LinearRegression()
```

Train/fit lm on the training data.

In [59]:
```
lm.fit(X_train,y_train)
```

Out[59]:
```
LinearRegression()
```

Print out the coefficients of the model

In [60]:
```
lm.coef_
```

Out[60]:
```
array([25.98154972, 38.59015875,  0.19040528, 61.27909654])
```

Predicting Test Data

Now that we have fit our model, let's evaluate its performance by predicting off the test values!

Use lm.predict() to predict off the X_test set of the data.

In [61]:
```
lm.predict(X_test)
```

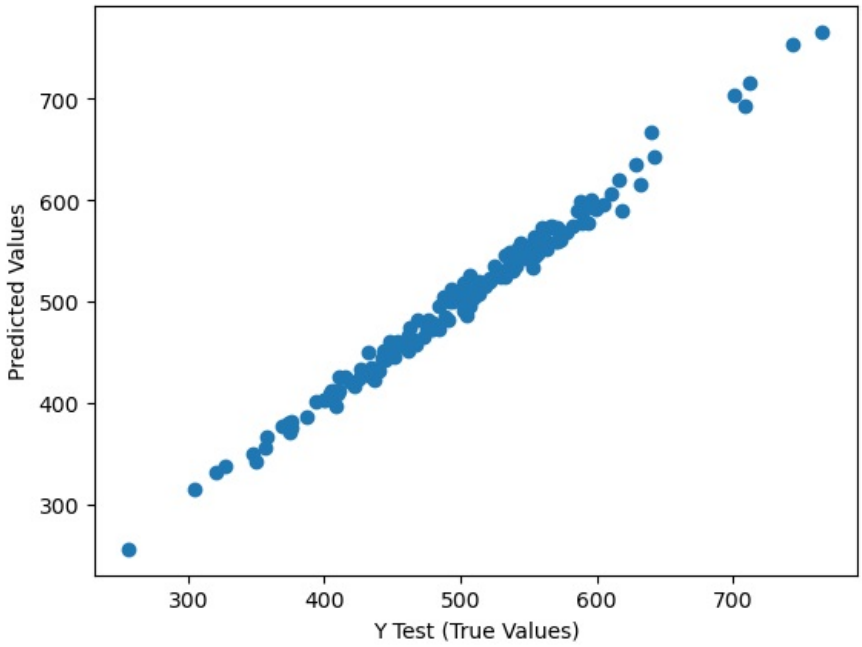
Out[61]:
```
array([456.44186104, 402.72005312, 409.2531539 , 591.4310343 ,
       590.01437275, 548.82396607, 577.59737969, 715.44428115,
       473.7893446 , 545.9211364 , 337.8580314 , 500.38506697,
       552.93478041, 409.6038964 , 765.52590754, 545.83973731,
       693.25969124, 507.32416226, 573.10533175, 573.2076631 ,
       397.44989709, 555.0985107 , 458.19868141, 482.66899911,
       559.2655959 , 413.00946082, 532.25727408, 377.65464817,
       535.0209653 , 447.80070905, 595.54339577, 667.14347072,
       511.96042791, 573.30433971, 505.02260887, 565.30254655,
       460.38785393, 449.74727868, 422.87193429, 456.55615271,
       598.10493696, 449.64517443, 615.34948995, 511.88078685,
       504.37568058, 515.95249276, 568.64597718, 551.61444684,
       356.5552241 , 464.9759817 , 481.66007708, 534.2220025 ,
       256.28674001, 505.30810714, 520.01844434, 315.0298707 ,
       501.98080155, 387.03842642, 472.97419543, 432.8704675 ,
       539.79082198, 590.03070739, 752.86997652, 558.27858232,
       523.71988382, 431.77690078, 425.38411902, 518.75571466,
       641.9667215 , 481.84855126, 549.69830187, 380.93738919,
       555.18178277, 403.43054276, 472.52458887, 501.82927633,
       473.5561656 , 456.76720365, 554.74980563, 702.96835044,
       534.68884588, 619.18843136, 500.11974127, 559.43899225,
       574.8730604 , 505.09183544, 529.9537559 , 479.20749452,
       424.78407899, 452.20986599, 525.74178343, 556.60674724,
       425.7142882 , 588.8473985 , 490.77053065, 562.56866231,
       495.75782933, 445.17937217, 456.64011682, 537.98437395,
       367.06451757, 421.12767301, 551.59651363, 528.26019754,
       493.47639211, 495.28105313, 519.81827269, 461.15666582,
       528.8711677 , 442.89818166, 543.20201646, 350.07871481,
       401.49148567, 606.87291134, 577.04816561, 524.50431281,
       554.11225704, 507.93347015, 505.35674292, 371.65146821,
       342.37232987, 634.43998975, 523.46931378, 532.7831345 ,
       574.59948331, 435.57455636, 599.92586678, 487.24017405,
       457.66383406, 425.25959495, 331.81731213, 443.70458331,
       563.47279005, 466.14764208, 463.51837671, 381.29445432,
       411.88795623, 473.48087683, 573.31745784, 417.55430913,
       543.50149858, 547.81091537, 547.62977348, 450.99057409,
       561.50896321, 478.30076589, 484.41029555, 457.59099941,
       411.52657592, 375.47900638])
```

In [62]:
```
#Let's call lm.predict - predictions

predictions = lm.predict(X_test)
```

Create a scatterplot of the real test values versus the predicted values.

In [63]:
```
plt.scatter(y_test,predictions)
plt.xlabel ('Y Test (True Values)')
plt.ylabel ('Predicted Values')

#The line is nearly perfect so there is few error between them
```

Out[63]:
```
Text(0, 0.5, 'Predicted Values')
```



Evaluating the Model

Let's evaluate our model performance by calculating the residual sum of squares and the explained variance score (R^2).

Calculate the Mean Absolute Error, Mean Squared Error, and the Root Mean Squared Error. Refer to the lecture or to Wikipedia for the formulas

In [64]:
```
from sklearn import metrics
```

In [66]:
```
print ('MAE' , metrics.mean_absolute_error(y_test,predictions))
print ('MSE' , metrics.mean_squared_error(y_test,predictions))
print ('RMSE' , np.sqrt(metrics.mean_squared_error(y_test,predictions)))
```

```
MAE 7.228148653430832
MSE 79.81305165097456
RMSE 8.93381506697864
```

In [67]:
```
#The folllowing is used to find our Explained variance score
metrics.explained_variance_score(y_test,predictions)
```

Out[67]:
```
0.9890771231889607
```

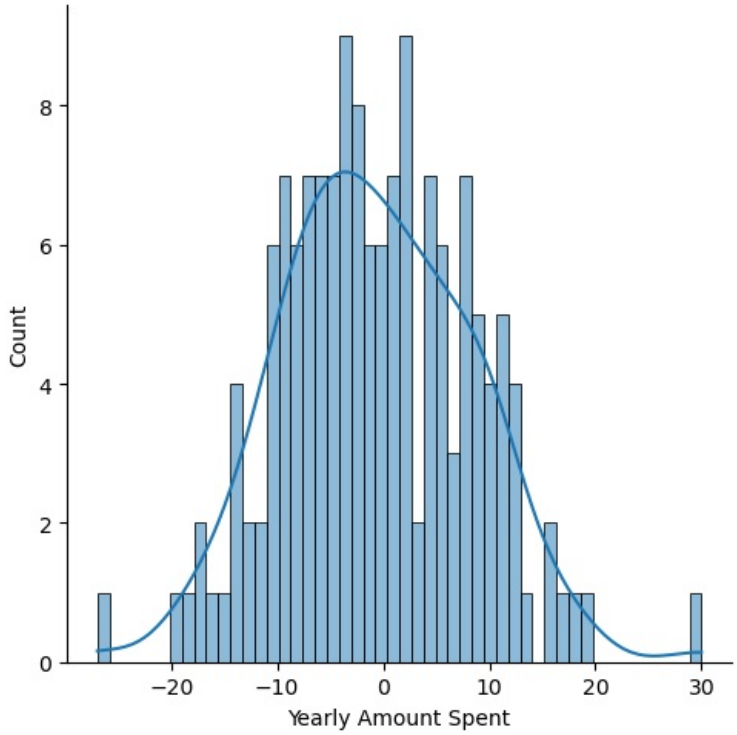
Residuals

You should have gotten a very good model with a good fit. Let's quickly explore the residuals to make sure everything was okay with our data.

Plot a histogram of the residuals and make sure it looks normally distributed. Use either seaborn distplot, or just plt.hist().

In [70]:
```
sns.displot((y_test-predictions),bins=50,kde=True)
```

Out[70]:
```
<seaborn.axisgrid.FacetGrid at 0x28876b273a0>
```



Conclusion

We still want to figure out the answer to the original question, do we focus our efforst on mobile app or website development? Or maybe that doesn't even really matter, and Membership Time is what is really important. Let's see if we can interpret the coefficients at all to get an idea.

Recreate the dataframe below.

In [72]:
```
cdf = pd.DataFrame(lm.coef_,X.columns,columns=['Coeff'])
cdf
```

Out[72]:

	Coeff
Avg. Session Length	25.981550
Time on App	38.590159
Time on Website	0.190405
Length of Membership	61.279097

How can you interpret these coefficients?

If we consider all things equal and fixed, an increase in 1 unit of a category like the Length of Membership (for example) will lead to an increase of $61.27 in Yearly Amount spent.

Do you think the company should focus more on their mobile app or on their website?

It depends on a lot of other external factors to be honest. But as a hypothesis, if the R&D component of the company knows that it can fastly and easily develop it's website component and depend on it more for operations, then they should focus more on it since it is cheaper and only brings about an increase of $0.2 on the yearly amount spent for an increase of 1 unit of time spent on developing the website. So at the end it also depends on what are the companies trade-offs and what are their goals for the long run.

Great Job!

Congrats on your contract work! The company loved the insights! Let's move on.