REPORT ON ARVILLA ECOMMERCE CLOTHING LINE CUSTOMERS EXPERIENCE USING MOBILE APP OR WEBSITE

Zep Internship Task

NOVEMBER 9, 2022

LINEAR REGRESSION PROJECT BY JERRY POLAND

NOTE: This was originally a capstone Machine learning project by Pierian Data (A Data science Training company)

1. PROBLEM STATEMENT

Arhilla is 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 an analyst on contract to help them figure it out!

The key objective of this analysis is to use Linear regression model, to develop a modelling framework that will tell whether the company should focus their efforts on their mobile app or their website.

2. THE DATASET

The dataset used is the Ecommerce Customers csv file from the company. It has Customer info, such as 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.

Lets have a view of the first five rows of the dataset

```
customers.head()
```

	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

Lets look at the description of the dataset

```
customers.describe()
```

	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

Let us as well consider the general information of the dataset

```
]: customers.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 8 columns):
Email                   500 non-null object
Address                 500 non-null object
Avatar                  500 non-null object
Avg. Session Length     500 non-null float64
Time on App             500 non-null float64
Time on Website         500 non-null float64
Length of Membership    500 non-null float64
Yearly Amount Spent     500 non-null float64
dtypes: float64(5), object(3)
memory usage: 31.3+ KB
```

3. DATA CLEANING

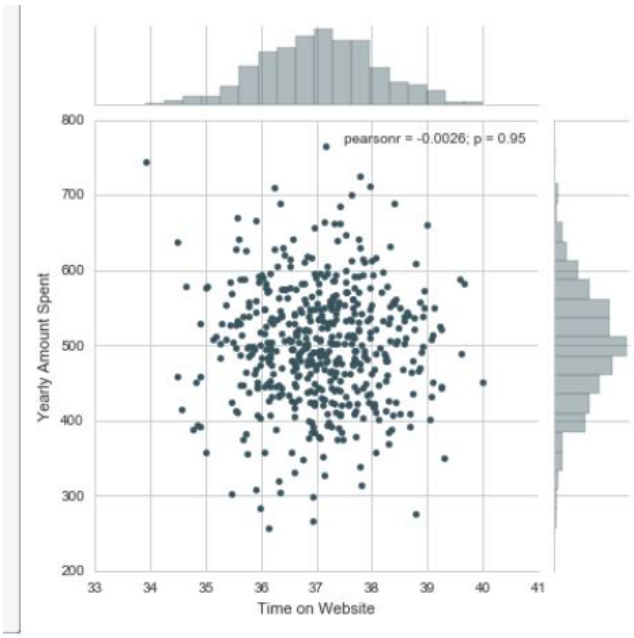
Data cleaning is a very important aspect in any data science project. The cleanness of our data determines the accuracy of our model. Looking at the 'customers.info()', we can tell the dataset is clean as there are no null values.

4. EXPLORATORY DATA ANALYSIS

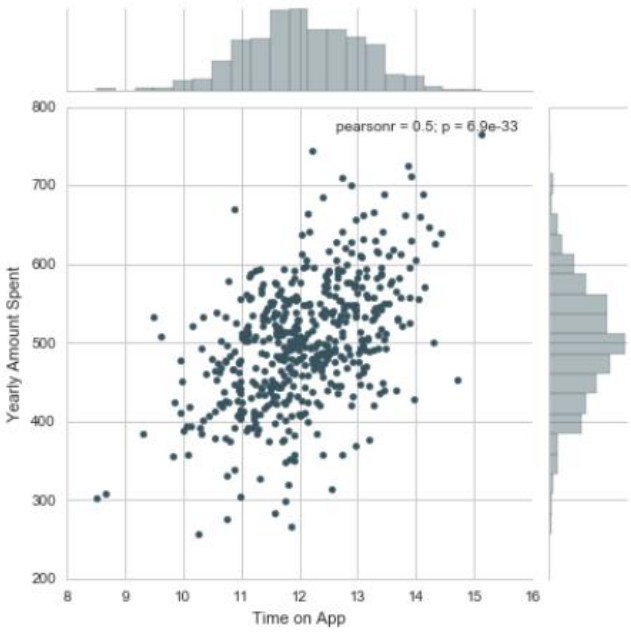
Exploratory Data Analysis is mandatory to explore in depth our data in other to draw meaningful insights from it. We will be using the numerical data of the csv file.

Correlation between columns.

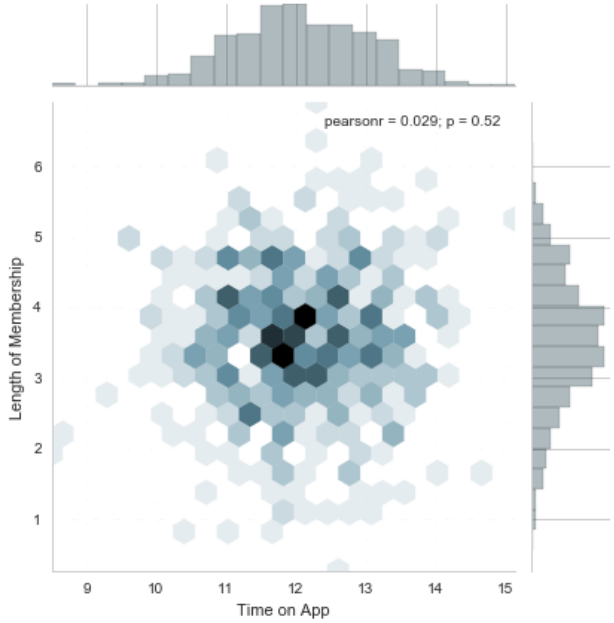
We will use seaborn to create a jointplot in order to make these comparisons.



Correlation between time on website and yearly amount spent columns

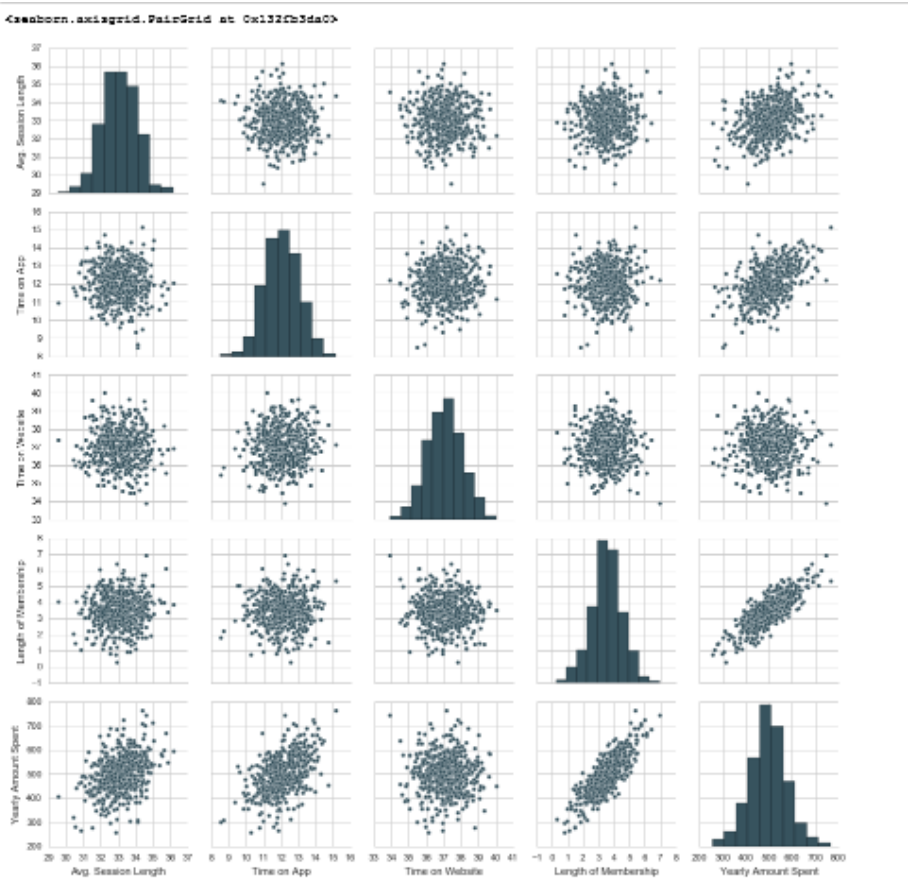


Correlation between time on app and yearly amount spent columns



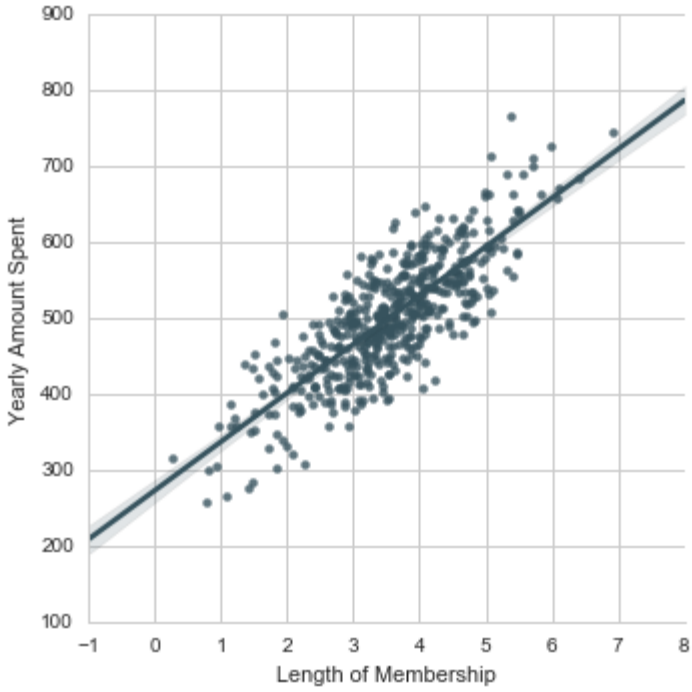
2D hex bin plot comparing Time on App and Length of Membership

Let's explore these types of relationships across the entire data set using pairplot



Based off the plot above what looks to be the most correlated feature with Yearly Amount Spent is the length of membership.

Thus we will create a linear model plot (using seaborn's lmplot) of Yearly Amount Spent vs. Length of Membership.



5. MODELING BUILDING

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.

Training and testing data

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

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

** 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**

```
from sklearn.model_selection import train_test_split
```

```
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 **

```
from sklearn.linear_model import LinearRegression
```

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

```
lm = LinearRegression()
```

** Train/fit lm on the training data.**

```
lm.fit(X_train,y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

Print out the coefficient of the model

```
# The coefficients
print('Coefficients: \n', lm.coef_)
```

```
Coefficients:
 [ 25.98154972  38.59015875   0.19040528  61.27909654]
```

Predicting Test Data

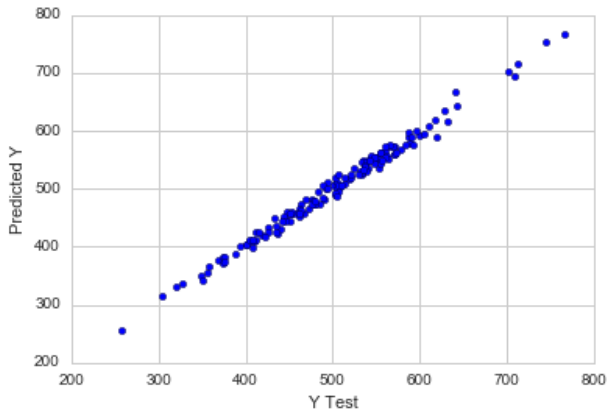
Now that we have fit our model, let's evaluate its performance by predicting off the test values!
** Using lm.predict() to predict off the X_test set of the data.**

```
predictions = lm.predict( X_test)
```

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

```
plt.scatter(y_test,predictions)
plt.xlabel('Y Test')
plt.ylabel('Predicted Y')
```

```
<matplotlib.text.Text at 0x135546320>
```



6. 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.**

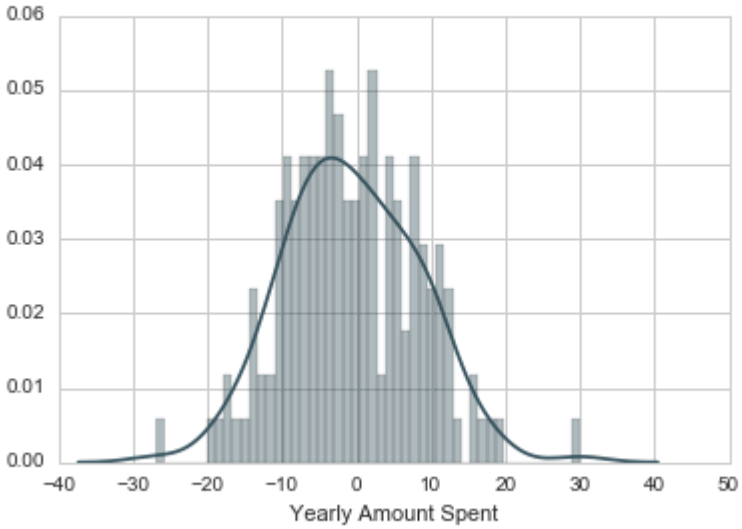
```
# calculate these metrics by hand!
from sklearn import metrics

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: 7.22814865343
MSE: 79.813051651
RMSE: 8.93381506698
```

Residual

Plot a histogram of the residuals and make sure it looks normally distributed. Use either seaborn distplot

```
sns.distplot((y_test-predictions),bins=50);
```



7. FINAL INTERPRETATION AND CONCLUSION

We still want to figure out the answer to the original question, do we focus our effort 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.

```
coeffecients = pd.DataFrame(lm.coef_,X.columns)
coeffecients.columns = ['Coeffecient']
coeffecients
```

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

Interpreting the coefficients:

- Holding all other features fixed, a 1 unit increase in **Avg. Session Length** is associated with an **increase of 25.98 total dollars spent**.
- Holding all other features fixed, a 1 unit increase in **Time on App** is associated with an **increase of 38.59 total dollars spent**.
- Holding all other features fixed, a 1 unit increase in **Time on Website** is associated with an **increase of 0.19 total dollars spent**.
- Holding all other features fixed, a 1 unit increase in **Length of Membership** is associated with an **increase of 61.27 total dollars spent**.

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

This is tricky, there are two ways to think about this: Develop the Website to catch up to the performance of the mobile app, or develop the app more since that is what is working better. This sort of answer really depends on the other factors going on at the company, we would probably want to explore the relationship between Length of Membership and the App or the Website before coming to a conclusion