KIBERNUM
IT Academy

Módulo 3 – Análisis Exploratorio y Programación Estadística

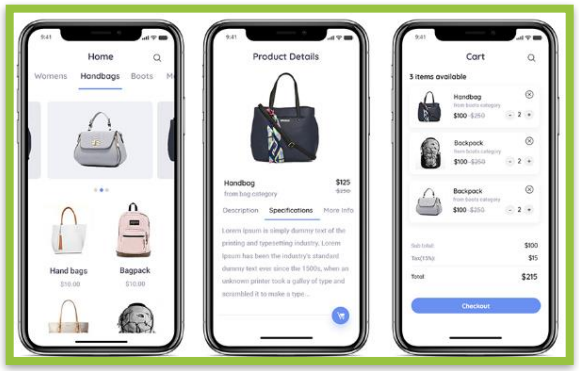
# Regresiones Lineales Múltiples

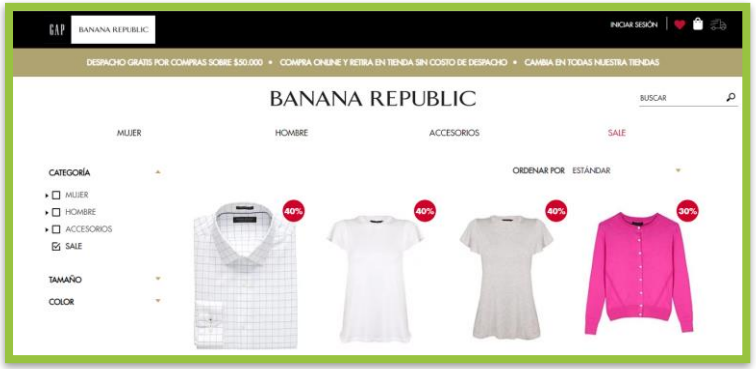
Ciencia de Datos

# Caso e-commerce

Usted acaba de ser contratado para trabajar en una compañía puntocom, ecommerce, establecida en la ciudad de New York dedicada a la venta de ropa exclusiva. La compañía realiza eventos con sus clientes con desfiles de moda y con asesores de estilo, quienes presentan los productos y aconsejan a los clientes. Esta tienda es exclusiva, por lo tanto, para participar de los eventos, los clientes deben tener una membresía.

Posteriormente, los clientes llegan a sus casas y realizan la compra, ya sea a través del website o del mobile app, de los productos que les fueron presentados y al día siguiente reciben las prendas en su domicilio.




KIBERNUM

# Caso e-commerce

En este momento, la compañía está tratando de decidir si enfoca sus esfuerzos en su mobile app o en su website. Es por esto que requieren de un analista de datos, para que analice la información y pueda realizar una recomendación al directorio. Para este propósito, se cuenta con un dataset que contiene la información de los clientes de la compañía y de su actividad en los canales de compra.

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

KIBERNUM

# Caso e-commerce

El dataset contiene la siguiente información relevante:

- **Avg. Session Length:** Tiempo promedio de las sesiones asesoramiento de estilo que se realizan en la tienda.
- **Time on App:** Tiempo promedio de permanencia en la App mobile en minutos.
- **Time on Website:** Tiempo promedio de permanencia en el sitio web en minutos.
- **Length of Membership:** Tiempo de membresía, es decir, cuántos años el cliente ha sido miembro.
- **Yearly Amount Spent:** Promedio de gasto en compras realizado de forma anual.

Predictores

| | 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 |

# Caso e-commerce

Predictores

| | 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 |

KIBERNUM

# Caso e-commerce

La pregunta que se quiere contestar es la siguiente:

- **¿Se puede hacer un modelo que explique el gasto anual que tienen los clientes en función de las variables descritas?**
- **A la luz del modelo, ¿qué conviene más? ¿Invertir en potenciar el website o la mobile app?**

KIBERNUM

# Regresiones Lineales Multivariadas

Las regresiones lineales simples son una aproximación útil para predecir una respuesta a partir de un único predictor.

| | | |
|---|---|---|
| Gasto Anual Cliente | ~ | Tiempo promedio permanencia sitio web |

Sin embargo, en la práctica a menudo hay más de un predictor que contribuye a explicar la respuesta de una variable de resultado.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Gasto Anual Cliente | ~ | Tiempo promedio permanencia sitio web | + | Tiempo promedio permanencia mobile app | + | Tiempo promedio permanencia sesiones asesoramiento | + | Tiempo de permanencia con la membresía |

¡¡ Recuerde que debe haber causalidad entre las variables predictoras y la variable outcome !!

KIBERNUM

# Regresiones Lineales Multivariadas

Regresión Lineal Simple

$$y = b_0 + b_1 x_1$$

Regresión Lineal Múltiple

$$y = b_0 + b_1 x_1 + \cdots + b_n x_n$$

Predictores

# Estimando los coeficientes

Al igual que en las regresiones lineales, los coeficientes $b_0, b_1, \ldots, b_n$ son desconocidos y por lo tanto deben ser estimados.

$$y = b_0 + b_1 x_1 + \cdots + b_n x_n$$

Dados los estimadores $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_n$ podemos hacer predicciones usando la fórmula:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \cdots + \hat{b}_n x_n$$

Los parámetros son estimados utilizando la aproximación de la suma de los errores cuadráticos, al igual que en la regresión lineal simple. Los valores de $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_n$ son aquellos que minimizan la suma del error cuadrático:

# Estimando los coeficientes

Los parámetros son estimados utilizando la aproximación de la suma de los errores cuadráticos, al igual que en la regresión lineal simple. Los valores de b^0,b^1…,b^n son aquellos que minimizan la suma del error cuadrático.

$$RSS = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

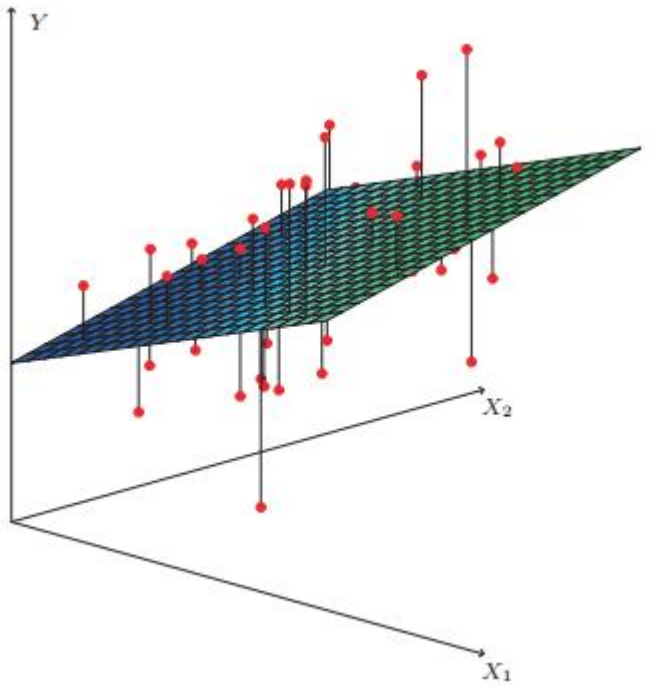
$$RSS = \sum_{i=1}^{m} (y_i - \hat{b}_0 - \hat{b}_1 x_1 - \cdots - \hat{b}_n x_n)^2$$

En donde,
m : cantidad de muestras
n : cantidad de predictores

KIBERNUM

# Variables

- En la siguiente figura se aprecia, para el caso de dos variables predictivas, el plano que minimiza la suma de los errores cuadráticos.




KIBERNUM

# Suposiciones de una Regresión Lineal

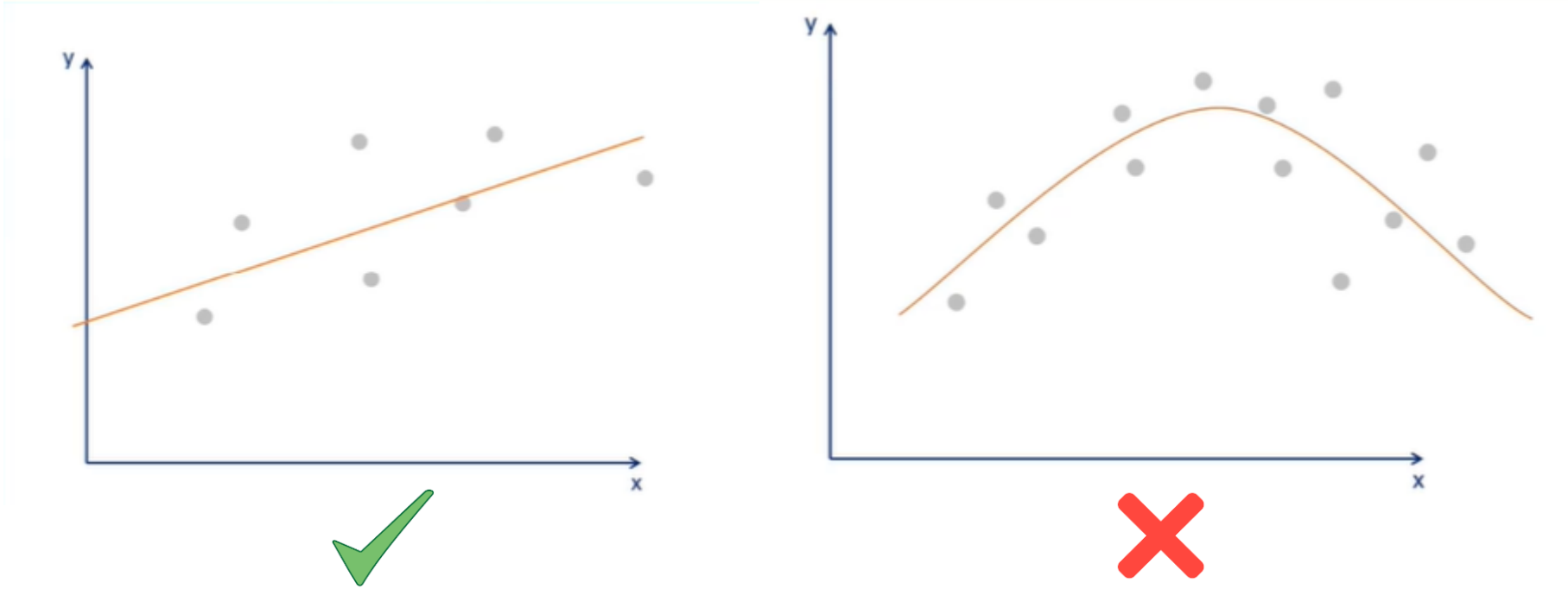
- Linearidad.
- No Endogeneidad.
- Homocedasticidad.
- No Autocorrelación.
- No Multicolinearidad.

KIBERNUM

# Linealidad

La regresión lineal asume que las variables independientes producen una respuesta lineal en la variable dependiente o resultado (outcome).

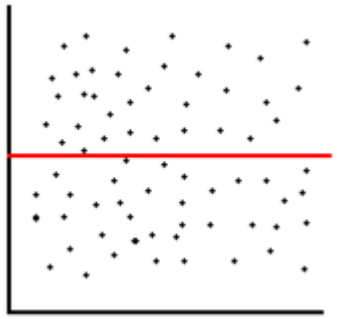
$$y = b_0 + b_1 x_1 + \cdots + b_n x_n$$

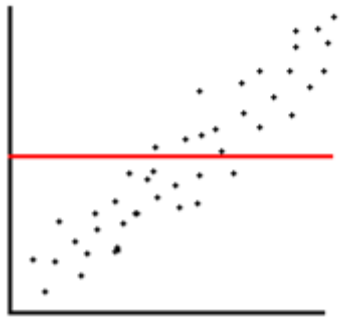


# No Endogeneidad

Se dice que hay endogeneidad cuando una variable independiente del modelo está correlacionada con el error. Esto puede suceder, por ejemplo, cuando falta incorporar una variable al modelo y dicha variable está correlacionada con la variable dependiente.

$$y = b_0 + b_1 x_1 + \varepsilon$$

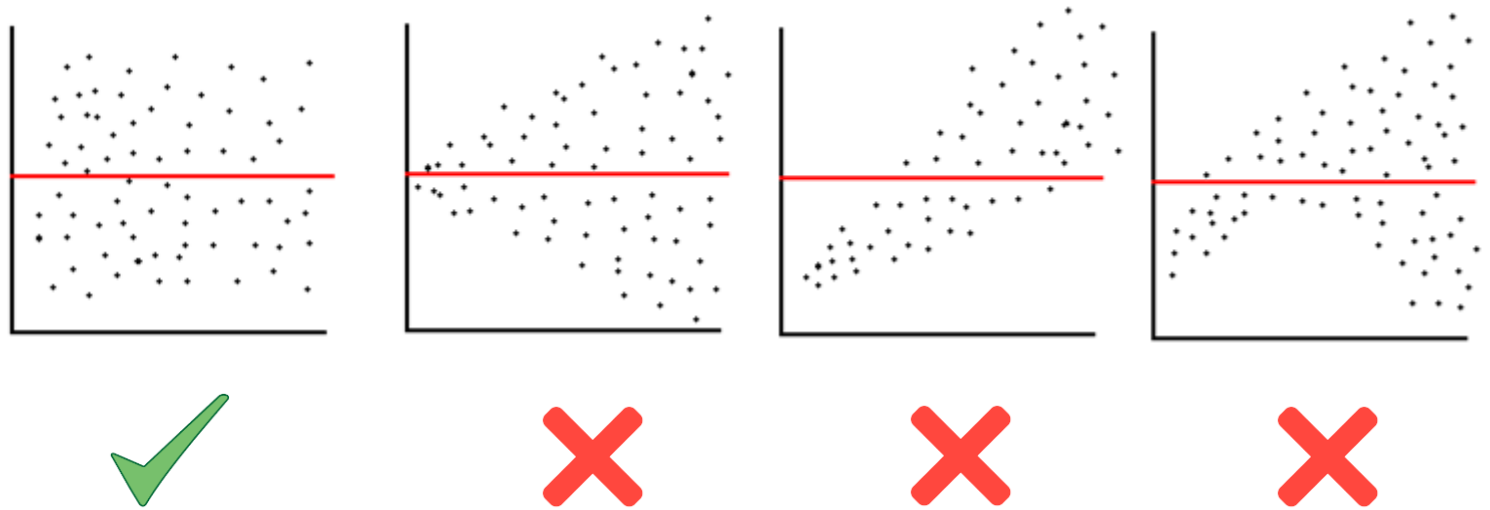




KIBERNUM

# Homocedasticidad

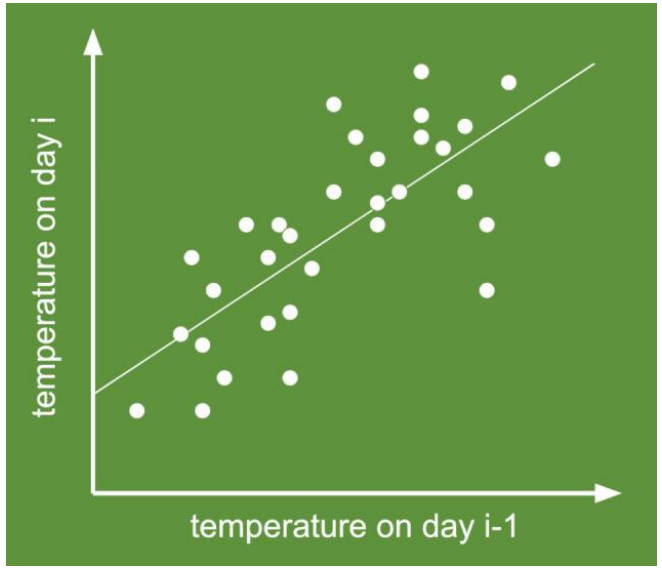
Se dice que un modelo predictivo presenta homocedasticidad cuando la varianza del error condicional a las variables explicativas es constante a lo largo de las observaciones.



KIBERNUM

# No Autocorrelación

La Autocorrelación se da cuando una variable está correlacionada consigo misma, por ejemplo, cuando una medición depende de la medición anterior. Esto se da a veces en las series de tiempo. Por ejemplo, para explicar la temperatura un día determinado, un predictor es la temperatura que hubo el día anterior.

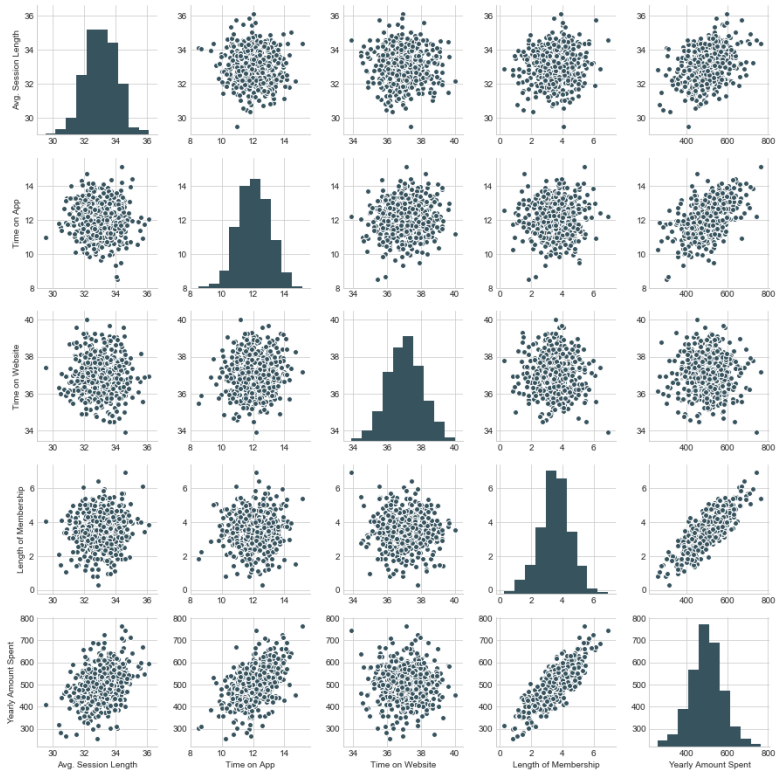


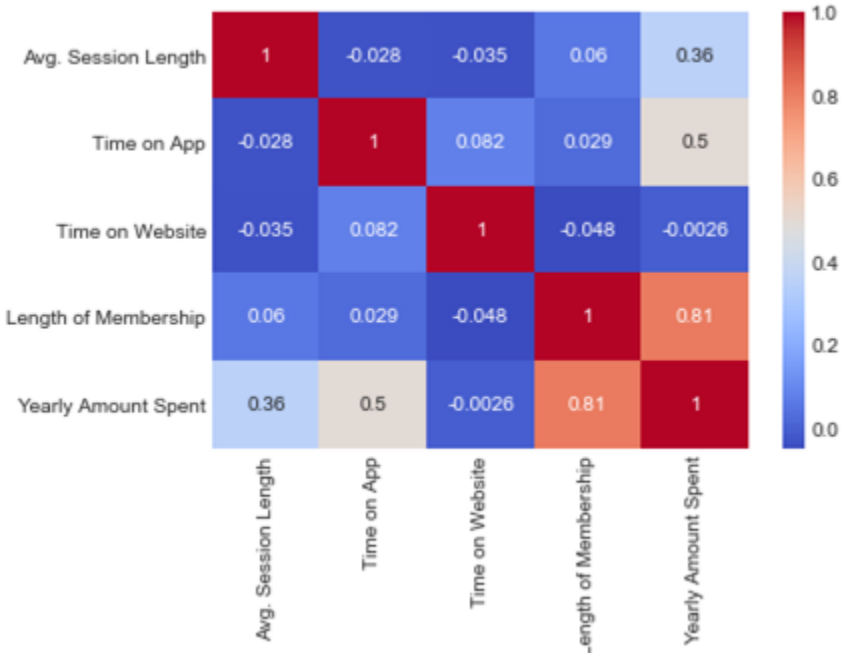

En ese caso, se recomienda utilizar otros tipos de modelos (p. ej: MA, ARMA, ARIMA)

KIBERNUM

# No Multicolinealidad

La multicolinealidad se da cuando hay variables que tienen una alta correlación entre sí. Una forma de detectarla es construyendo una matriz de correlación entre las variables. Si una variable tiene un valor alto, por ejemplo, sobre 0.7, presenta una fuerte correlación y debería considerarse su eliminación del modelo.



# Implementación en Python

KIBERNUM

# Análisis de Correlación

Un **supuesto de los modelos de regresión lineal múltiple es la multicolinealidad. Es decir, las variables predictoras no deben estar correlacionadas.**

```
clientes.corr()
```

| | Avg. Session Length | Time on App | Time on Website | Length of Membership | Yearly Amount Spent |
|---|---|---|---|---|---|
| **Avg. Session Length** | 1.000000 | -0.027826 | -0.034987 | 0.060247 | 0.355088 |
| **Time on App** | -0.027826 | 1.000000 | 0.082388 | 0.029143 | 0.499328 |
| **Time on Website** | -0.034987 | 0.082388 | 1.000000 | -0.047582 | -0.002641 |
| **Length of Membership** | 0.060247 | 0.029143 | -0.047582 | 1.000000 | 0.809084 |
| **Yearly Amount Spent** | 0.355088 | 0.499328 | -0.002641 | 0.809084 | 1.000000 |

Como se puede observar, las variables predictoras poseen una correlación cercana a cero, por lo tanto, se puede afirmar que son variables independientes entre sí.

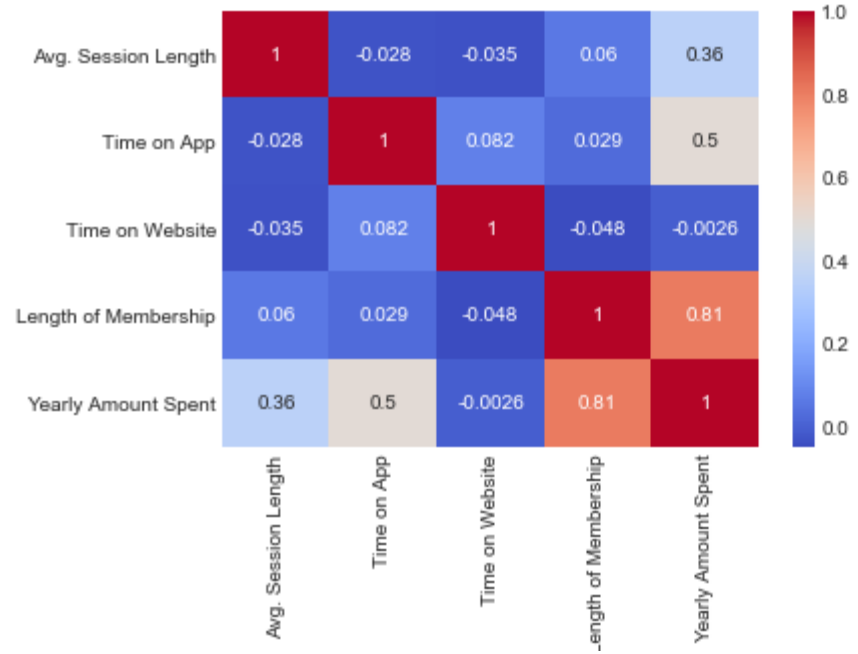
KIBERNUM

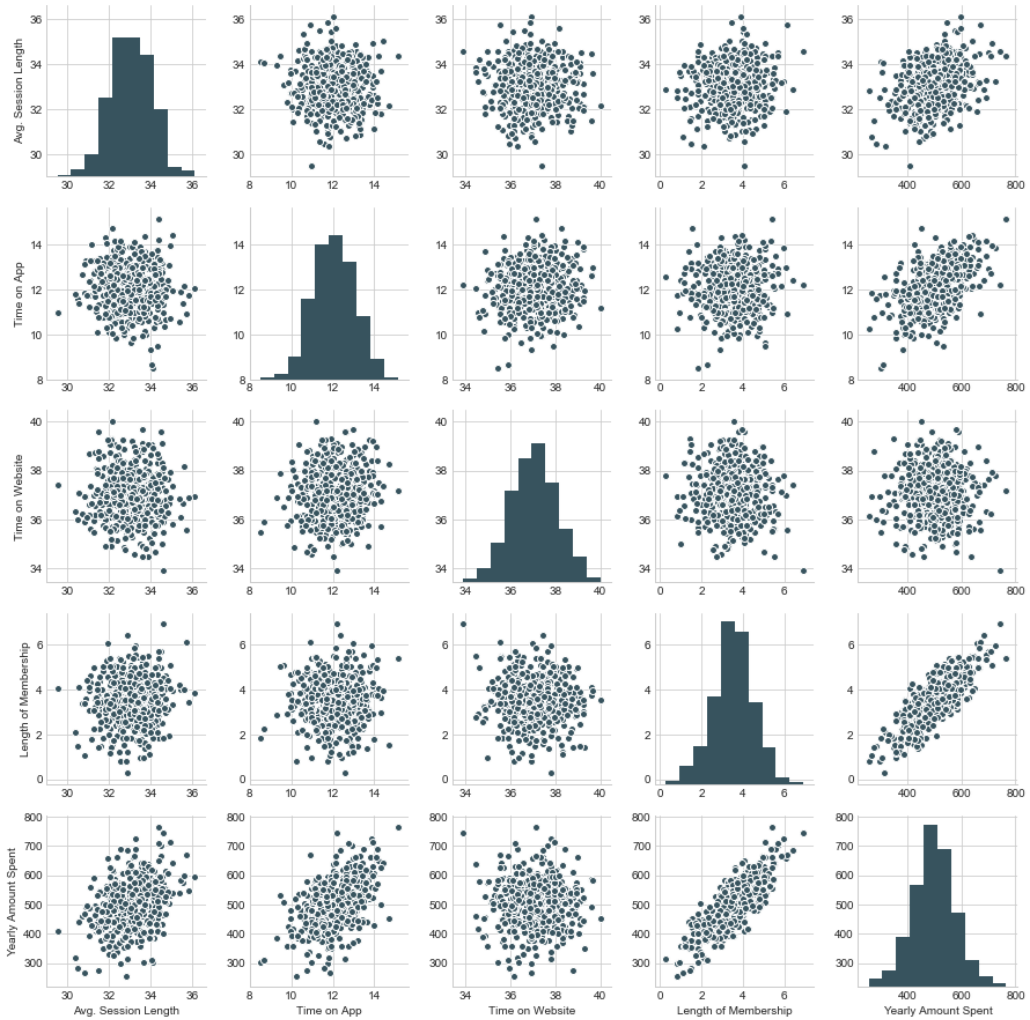
# Análisis de Correlación

El siguiente mapa de calor permite visualizar de forma más fácil las correlaciones entre variables. Nótese que hay una fuerte correlación entre el predictor Length of Membership y la predicción Yearly Amount Spent.

```
sns.heatmap(clientes.corr(), cmap="coolwarm", annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2b60df101d0>
```






# Análisis de Correlación

Lo mismo se puede apreciar en el siguiente **joinplot :**

```
sns.set_palette("GnBu_d")
sns.set_style('whitegrid')
sns.pairplot(clientes)
```

KIBERNUM

# Formulación del Modelo

Como se puede observar, la fórmula del modelo considera todos los predictores disponibles en el set de datos.

Si queremos escribir la fórmula en varias líneas, debemos indicarlo con triple comilla al inicio y al fin.

```
lm = smf.ols(formula='''Q("Yearly Amount Spent")
                        ~ Q("Avg. Session Length") +
                          Q("Time on App") +
                          Q("Time on Website") +
                          Q("Length of Membership")''',data=clientes).fit()
```

KIBERNUM

# Resultado del Entrenamiento

A continuación, revisaremos los valores de los coeficientes y del intercepto que se han ajustado en el modelo.

```
lm.params
```

```
Intercept                   -1051.594255
Q("Avg. Session Length")       25.734271
Q("Time on App")               38.709154
Q("Time on Website")            0.436739
Q("Length of Membership")      61.577324
dtype: float64
```

Estos valores se interpretan de la siguiente manera:

Por ejemplo, el aumento en 1 unidad (es decir, 1 minuto) de Avg. Session Length está asociado con el aumento en 25.98 dólares en Yearly Amount Spent. Lo mismo para las demás variables.

KIBERNUM

```
lm.summary()
```

| | | | |
|---|---|---|---|
| Model: | OLS | Adj. R-squared: | 0.984 |
| Dependent Variable: | Q("Yearly Amount Spent") | AIC: | 3723.8197 |
| Date: | 2020-06-18 12:37 | BIC: | 3744.8927 |
| No. Observations: | 500 | Log-Likelihood: | -1856.9 |
| Df Model: | 4 | F-statistic: | 7766. |
| Df Residuals: | 495 | Prob (F-statistic): | 0.00 |
| R-squared: | 0.984 | Scale: | 99.466 |

| | Coef. | Std.Err. | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | -1051.5943 | 22.9925 | -45.7363 | 0.0000 | -1096.7693 | -1006.4193 |
| **Q("Avg. Session Length")** | 25.7343 | 0.4510 | 57.0571 | 0.0000 | 24.8481 | 26.6204 |
| **Q("Time on App")** | 38.7092 | 0.4510 | 85.8282 | 0.0000 | 37.8230 | 39.5953 |
| **Q("Time on Website")** | 0.4367 | 0.4441 | 0.9834 | 0.3259 | -0.4358 | 1.3093 |
| **Q("Length of Membership")** | 61.5773 | 0.4483 | 137.3463 | 0.0000 | 60.6964 | 62.4582 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.337 | Durbin-Watson: | 1.887 |
| Prob(Omnibus): | 0.845 | Jarque-Bera (JB): | 0.198 |
| Skew: | -0.026 | Prob(JB): | 0.906 |
| Kurtosis: | 3.083 | Condition No.: | 2642 |

# Resultado del Entrenamiento

Acá obtenemos el sumario del modelo ajustado.

KIBERNUM

# Métricas de evaluación

Calculamos las métricas de error del modelo con todas las variables.

```
y_true = clientes['Yearly Amount Spent']
y_pred = lm.predict(clientes[['Avg. Session Length','Time on App','Time on Website','Length of Membership']])
print( 'MAE: {}'.format(metrics.meanabs(y_true,y_pred)) )
print( 'MSE: {}'.format(metrics.mse(y_true,y_pred)) )
print( 'RMSE: {}'.format(metrics.rmse(y_true,y_pred)) )
print( 'R2: {}'.format(lm.rsquared))
print( 'R2-Adj: {}'.format(lm.rsquared_adj))
```

```
MAE: 7.877162860953783
MSE: 98.47102522149004
RMSE: 9.923256785022247
R2: 0.9843155370226726
R2-Adj: 0.9841887938875022
```

En este ejemplo, el MAE indica que el modelo, en promedio, tiene un error de 7.87 dólares. Por otra parte, el RMSE indica que, en promedio, el modelo tiene un error de 9,92 dólares en sus predicciones.

# Realizando predicciones

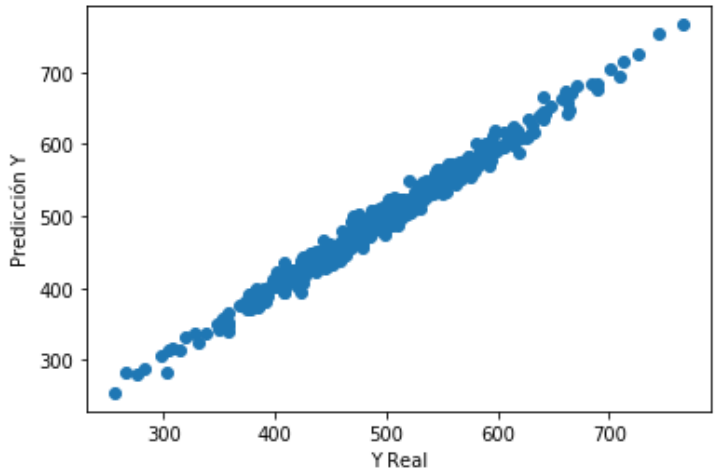
A continuación, vamos a realizar predicciones en el set de datos de Test y lo vamos a comparar con los valores reales de y.

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

En el siguiente gráfico comparamos los valores predichos versus los valores reales de “y” en el set de test. Como se puede apreciar, el modelo es bastante certero.

```
plt.scatter(y_true,y_pred)
plt.xlabel('Y Real')
plt.ylabel('Predicción Y')
```

```
Text(0, 0.5, 'Predicción Y')
```



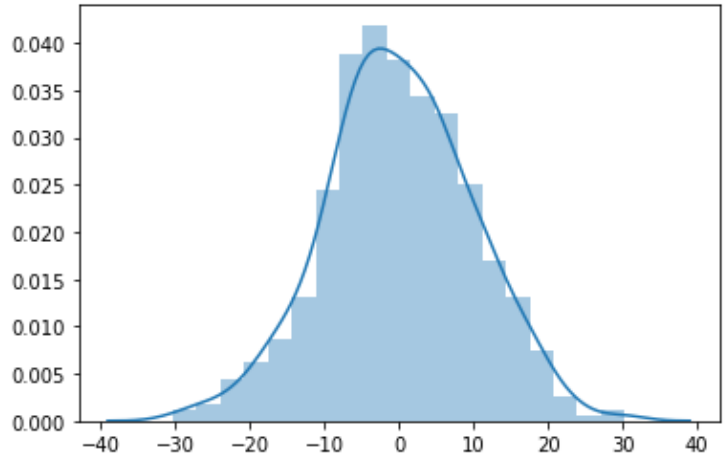

KIBERNUM

# Análisis de Residuales

Por último, verificamos que efectivamente los residuales se distribuyen de forma normal en el modelo.

```
sns.distplot(y_true-y_pred)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x2b6a6cd6b48>
```




KIBERNUM

# Dudas y consultas

KIBERNUM

Fin Presentación

KIBERNUM