Regresie liniară

Unul din principalele capitole ale statisticii are în vedere posibilitatea de a face predicții. Prin intermediul regresiei se pot face predicții ale unei variabile, în funcție de valoarea alteia. Predicția este procesul de estimare a valorii unei variabile cunoscând valoarea unei alte variabile.

În continuare, ne vom referi doar la situația regresiei simple (o variabilă dependentă și una independentă) și liniare (relația dintre cele doua variabile poate fi descrisă printr-o dreaptă în cadrul norului de puncte).

Regresia se leagă foarte mult de conceptul de corelație. O asociere puternică între două elemente conduce la creșterea preciziei predicției unei variabile pe seama alteia.

Exemplu de proiect pentru crearea unui model liniar:

O companie de comerț electronic cu sediul în New York City, vinde îmbrăcăminte online, dar au și sesiuni de consiliere în materie de stil și îmbrăcăminte în magazin. Clienții vin în magazin, au ședințe/întâlniri cu un stilist personal, apoi pot merge acasă și pot comanda fie pe o aplicație mobilă, fie pe site pentru hainele pe care le doresc.

Compania încearcă să decidă dacă își concentrează eforturile pe experiența în aplicația mobilă sau pe site-ul lor web. Te-au angajat cu contract pentru a-i ajuta să-și dea seama! Să începem!

Doar urmați pașii de mai jos pentru a analiza datele clienților (este fals, nu vă faceți griji, nu v-am dat numere reale de card de credit sau e-mailuri).

1. Importarea datelor

Pasul 1: Importați panda, numpy, matplotlib și seaborn. Apoi setați %matplotlib inline (veți importa apoi sklearn după necesitate).

2. Obținerea datelor

Vom lucra cu fișierul Ecommerce Customers CSV de la companie. Are informații despre clienți, cum ar fi e-mail, adresă și avatar de culoare. Apoi are și coloane cu valori numerice:

- Avg. Session Length: durata medie a sesiunilor de consiliere stil în magazin.
- Time on App: timpul mediu petrecut pe aplicație în minute.
- Time on Website: timpul mediu petrecut pe site în minute
- Length of Membership: de câți ani este membru clientul.

Pasul 2: Citiți fișierul Ecommerce Customers CSV ca DataFrame numit customers.

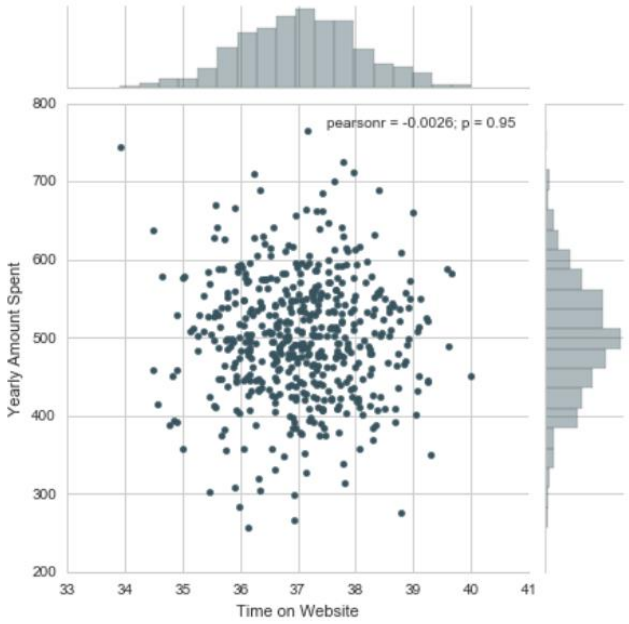
Pasul 3: Citiți primele linii (head) despre clienți și aflați informații despre ei folosinf funcțiile info() și describe().

3. Analiza exploratorie a datelor

Pentru restul exercițiului vom folosi doar datele numerice ale fișierului csv.

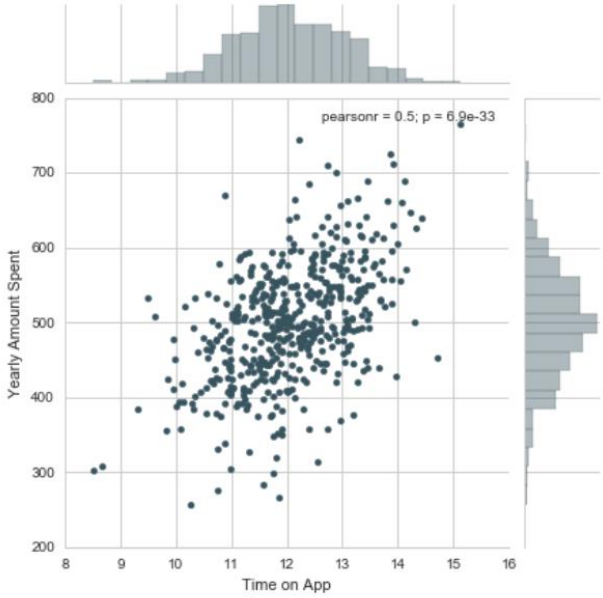
Utilizați seaborn pentru a crea o diagramă comună (**jointplot**) pentru a compara coloanele **Time on Website** și **Yearly Amount Spent**. Are sens corelația?

```
<seaborn.axisgrid.JointGrid at 0x120bfcc88>
```

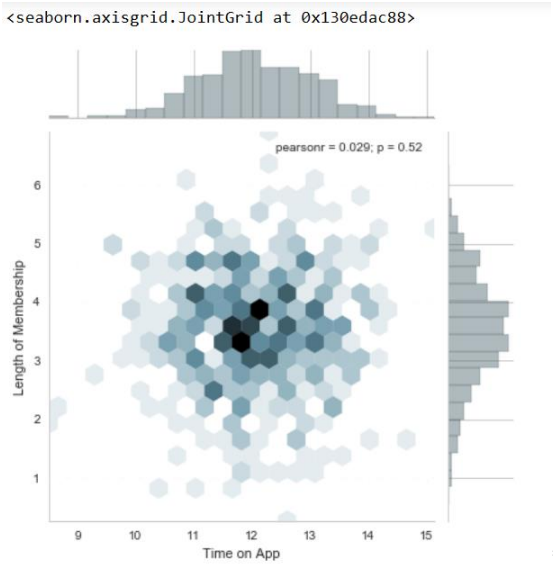


Pasul 4: Faceți același lucru, dar cu coloana Time on App.

```
<seaborn.axisgrid.JointGrid at 0x132db5908>
```

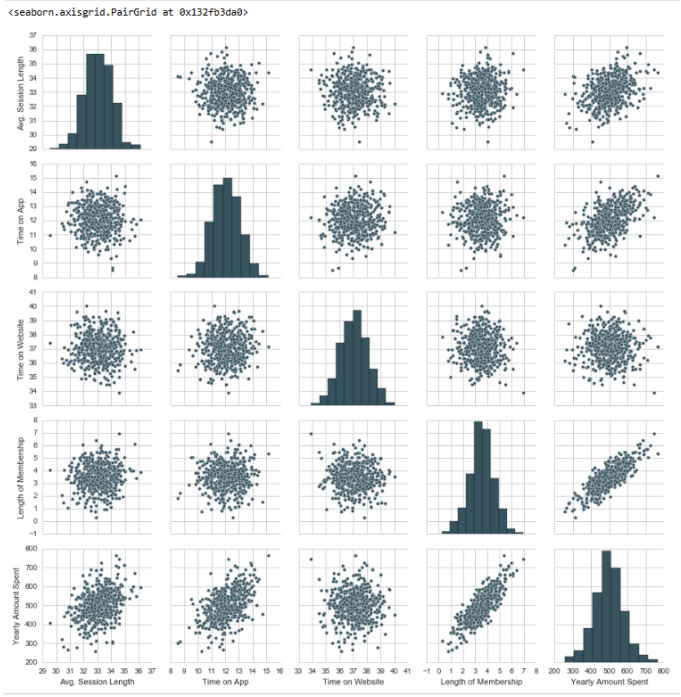


Pasul 5: Folosiți jointplot() pentru a crea o diagramă 2D ce compară Time on App și Length of Membership.



*(kind='hex')

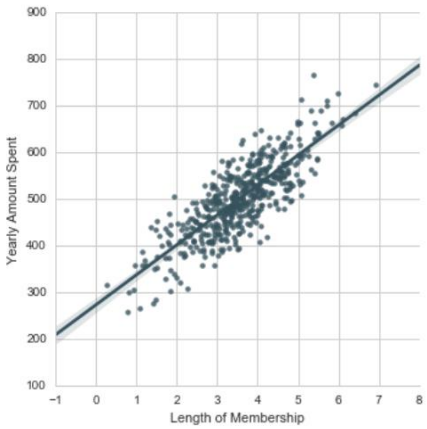
Pasul 6: Să explorăm aceste tipuri de relații în întregul set de date. Utilizați pairplot pentru a recrea graficul de mai jos. (Nu vă faceți griji pentru culori)



Pe baza acestei diagrame, care pare a fi cea mai corelată caracteristică cu Yearly Amount Spent?

Pasul 7: Creați un model liniar grafic (folosind Implot din seaborn) al datelor: Amount Spent vs. Length of Membership.

```
<seaborn.axisgrid.FacetGrid at 0x13538d0b8>
```



4. **Date de instruire și testare**

Să împărțim datele în seturi de antrenament și de testare.

Pasul 8: Setați o variabilă X egală cu caracteristicile numerice ale clienților și o variabilă y egală cu coloana Yearly Amount Spent.

Pasul 9: Utilizați model_selection.train_test_split din sklearn pentru a împărți datele în seturi de antrenament și de testare. Setați test_size=0,3 și random_state=101.

5. **Antrenarea modelului**

Acum este timpul să ne instruim modelul pe datele noastre de antrenament!

Pasul 10: Importați LinearRegression din sklearn.linear_model.

Pasul 11: Creați o instanță a unui model LinearRegression() numit lm.

Pasul 12: Antrenați lm pe datele de antrenament. `LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)`

Pasul 13: Imprimați coeficienții modelului.

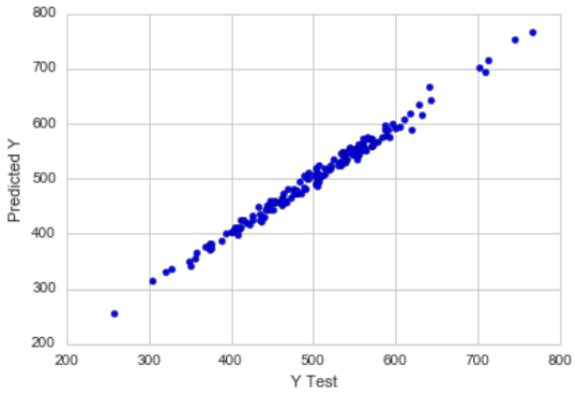
6. **Prezicerea datelor de testare**

Să evaluăm performanța modelului prin prezicerea valorilor testului!

Pasul 14: Utilizați lm.predict() pentru a prezice setul X_test de date.

Pasul 15: Creați o diagramă de dispersie a valorilor reale de test față de valorile prezise.

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



7. Evaluarea modelului

Să evaluăm performanța modelului nostru calculând suma reziduală a pătratelor și scorul de vari anță (R^2).

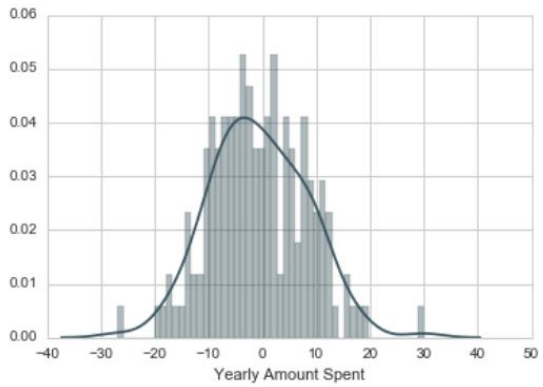
Pasul 16: Calculați eroarea medie absolută, eroarea medie pătratică și eroarea medie pătratică. Consultați prelegerea sau Wikipedia pentru formule.

```
MAE: 7.22814865343
MSE: 79.813051651
RMSE: 8.93381506698
```

8. Reziduuri

Să explorăm rapid reziduurile pentru a ne asigura că totul a fost în regulă cu datele noastre.

Pasul 17: Trasați o histogramă a reziduurilor și asigurați-te că are o distribuție normală. Utilizați fie seaborn distplot, fie doar plt.hist().



9. Concluzie

Încă vrem să aflăm răspunsul la întrebarea inițială, ne concentrăm eforturile pe dezvoltarea de aplicații mobile sau site-uri web? Sau poate că asta nici măcar nu contează, iar timpul de membru este ceea ce este cu adevărat important. Să vedem dacă putem interpreta coeficienții pentru a ne face o idee.

Pasul 18: Recreați setul de date de mai jos.

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

Pasul 19: Răspundeți la întrebare: Cum puteți interpreta acești coeficienți?
Pasul 20: Răspundeți la întrebare: Credeți că compania ar trebui să se concentreze mai mult pe aplicația mobilă sau pe site-ul lor web?

Sarcină individuală: Consultați https://www.kaggle.com/datasets pentru a alege și a descărca un set de date în scopul creării unui model liniar.
Dacă este necesar curățați mai întâi datele și argumentați metoda aleasă.
Parcurgeți pașii 1-20 ca în exemplu de mai sus - argumentați corelația dintre coloanele setului de date ales și evaluați predicția.