Visuell Dataanalys

Lektion X

2021-12-04

1 Seaborn och Pandas

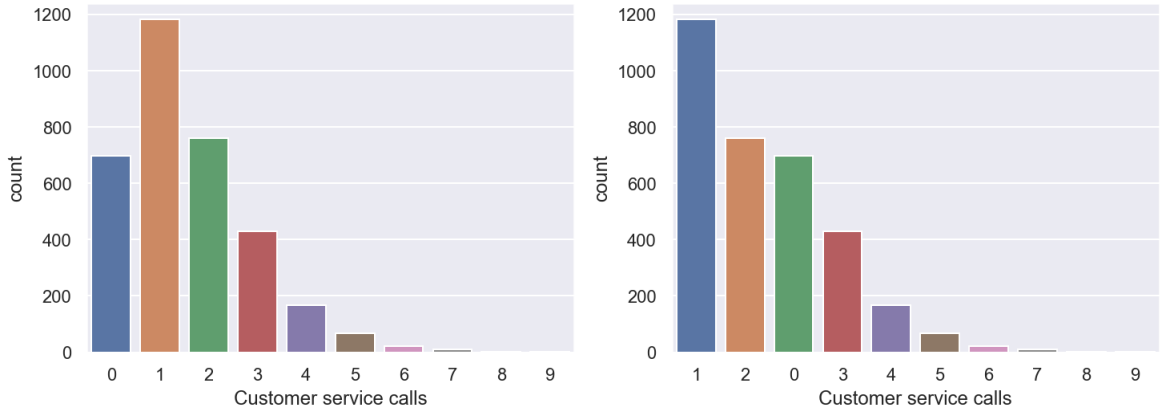
Ladda in Telecom Churn och inspektera vilken typ utav data det rör sig om samt antal dimensioner. Kan någon form utav encoding behöva göras?

Vår targetvariabel är 'Churn' som avgör huruvida en kund lämnar oss eller stannar. Klassimbalans för klassificering kan vara ett problem - avgör därför om datasetet är balanserat eller ej. (Är du intresserad av att veta mer om att åtgärda detta kan du googla SMOTE, data augmentation, over-/under sampling m.m.)!

Gör två stapeldiagram/bar plots för Total Day Minutes och Total intl calls. Detta är diskreta distributioner - testa att ange keyword kind = 'density' i plotten. Inspektera kurvorna och avgör om det finns någon logisk begränsning sett till vilka värden som kan antas för densitetskurvorna.

Testa att använda seaborns distplot, boxplot och violinplot för total intl calls. Vad ser du?

Använd seaborns countplot för att göra två bar plots med subplot av Customer service calls - en sorterad på index och en sorterad på count av index (som nedan)



Utvärdera hur stark korrelation det är mellan alla olika variabler i en heatmap. Finns det några variabler som är starkt korrelerade?

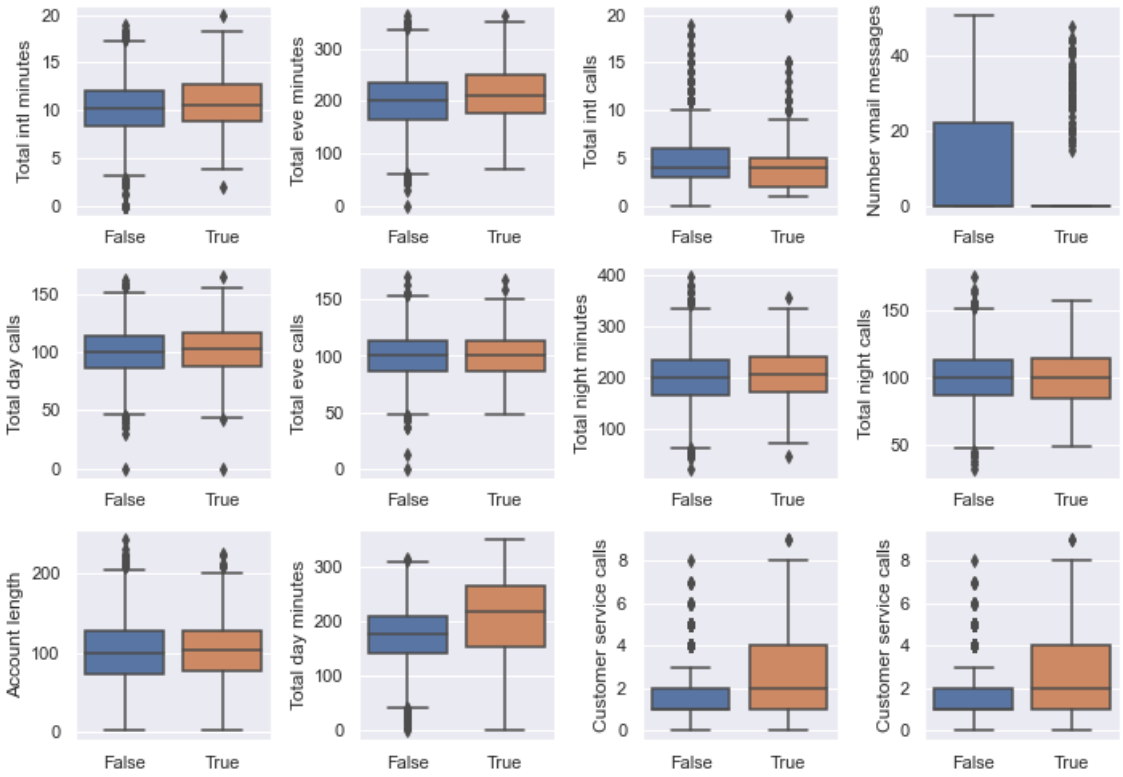
Testa att använda en scatter plot mellan två variabler som har låg korrelation och mellan två variabler som har hög korrelation. Hur skiljer sig formerna ut? Varför?

Testa att använda jointplot från Seaborn med kind = 'scatter' och kind = 'kde'. Hur skiljer de sig åt och hur tolkar du resultaten?

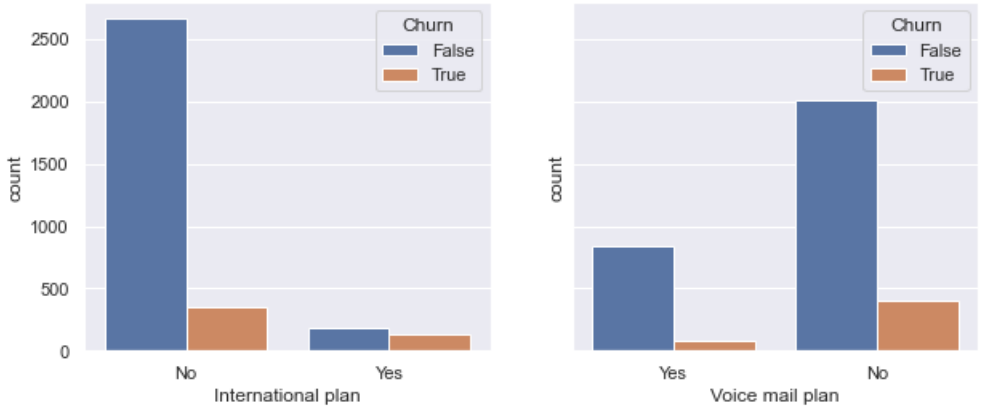
Använd paiplot från seaborn för alla numeriska variabler i datan. Hur tolkar du resultaten? Testa även att använda parametern 'hue' för att ge färgkodning baserat på Churn.

Använd lmplot med keyword hue = 'Churn' i scatter plots för valfria variabler. Testa högt korrelerade och lågt korrelerade variabler. Hittar du några samband?

Reproducera visualen nedan - notera att det är subplots. Här är alltså färgkodade distributioner för alla numeriska parametrar boxplottade. Hur skiljer denna vy sig mot pairplotten med färgkodning? Finns det för- och nackdelar med respektive? Hur skiljer sig beräkningstiden åt?



Reproducera, analysera och kommentera visualen under (notera att subplotsen delar y-axel.)



Sortera Churn rate för varje delstat från din ursprungliga dataframe - försök att göra det med en rad kod om du kan!

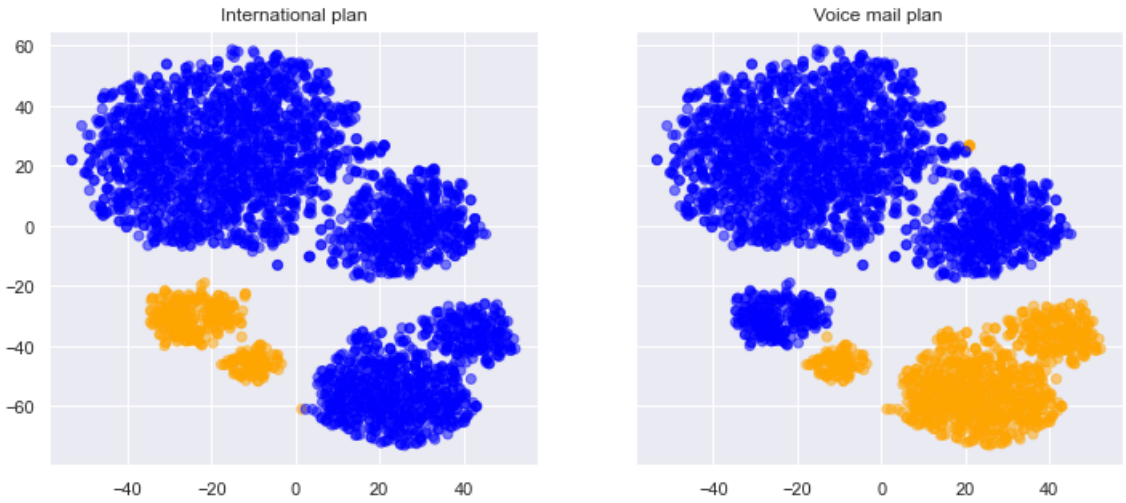
2 Dimensionsreducering

Läs på om dimensionsreduceringstekniken t-SNE och använd sklearns implementation av t-SNE för att dimensionsreducera din data. Gör först följande operationer:

- Plocka bort 'Churn' och 'State' - Churn behöver dock användas till färgkodning senare
- Encoda alla binära stringfeatures

Skala datan i dataframen och scatter plotta din dimensiosreducerade data. Färgkoda med avseende på Churn.

I en ny plot, scatter plotta och färgkoda Churn i två olika plots - en för International Plan och en för Voice Mail Plan. Vad säger de olika klustrena dig? Exempelbild på hur det kan se ut nedan.



Som för många andra saker här i livet finns det olika nivåer av förståelse - t-SNE är inget undantag. Att förstå hur den fungerar på djupet kräver en del matematisk förkunskap, men det är ingenting som krävs för att kunna implementera t-SNE eller tolka resultatet.

3 Extramaterial

Testa att ladda in MNIST Handwritten Digits och använd en ökande andel av datan för att dimensionsreducera. Hur beror beräkningstiden av mängden data? Ökar den linjärt? Om nej, ökar beräkningstiden långsammare eller snabbare?

Mixtra runt med parametern perplexity och sätt inget random seed. Får du samma resultat varje gång? Om nej, vad säger det dig?