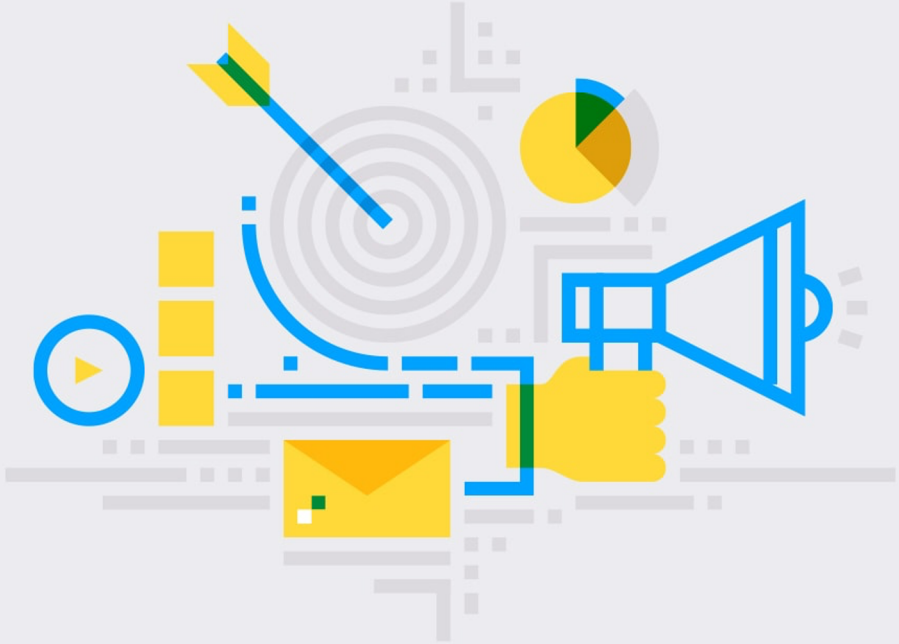
PERAKENDE PAZARLAMA İÇİN GERÇEKÇİ PAZAR SEPETİ VERİ SİMÜLASYONU

Merve Poslu

Büyük veriyi depolayabilen, işleyip zenginleştirebilen ve günlük hayatında kullanabilen perakendecilerin operasyonel verimliliğini ve müşteri memnuniyetini önemli derecede artırmaktadır. Perakendecilerin yetkinlikleri büyük veriye yönelik geliştirilmekte olan bilgi teknolojisi altyapı ve çözümleri sayesinde üç ana alanda artmaktadır:

- **1. Müşteriler hakkında hemen hemen herşeyi öğrenme yetkinliğimizin artması.**
- **2. Gerçek zamanlı veri toplanarak anlık verilmesi gereken kararların ve aksiyonların alınabilmesi.**
- **3. Hız ve ölçekte inovasyon sağlanması.**

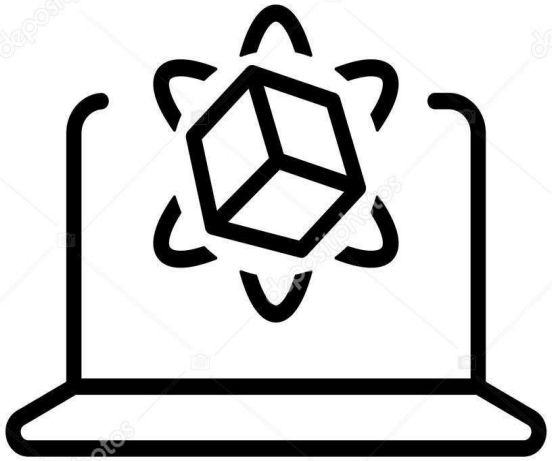


Bu Çalışma Neler İçerir?

- Perakendeciler, müşterilerinin çeşitli durumlarda satın alma alışkanlıklarını daha iyi anlayarak hizmet ve iş performansını artırabilir ve simülasyonlar, karşılaştırma için bir kıyaslama olarak kullanılabilir.
- Projenin amacı, satış tahmini ve iş senaryosu oluşturmada potansiyel kullanım için üç simülatörün etkinliğini değerlendirmektir.
- Her işlemin satın alınan ürünlerden oluşan bir "sepet" olarak kabul edildiği büyük bir satış noktası veri kümesi olan *The Complete Journey - Shopping Transaction Dataset* kullanılmıştır.
- Veri seti, bir perakendecide sık alışveriş yapan 2.500 haneden oluşan bir gruptan iki yıl boyunca hane düzeyindeki işlemleri içerir.

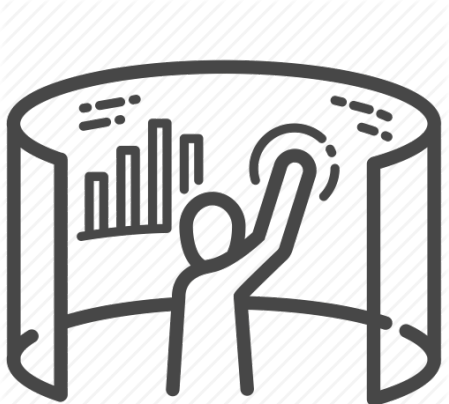
Simülasyon

- Veriler iki kısma ayrılmıştır: performans değerlendirmesi için holdout veriseti olarak rastgele seçilen bir haftanın verileri ve kalan kısım simülasyon modellemesi için veri modellerini belirlemeye yönelik eğitim seti olarak.
- Daha sonra haftalık sepetlerin simülasyonu yapılmıştır. Simüle edilmiş sepetler, gerçeğe ne kadar yakın olduklarını görmek için gerçek sepetlerle karşılaştırılır.
- Bu karşılaştırmalara dayanarak, perakende verilerini simüle etmek için en iyi modeli doğru bir şekilde belirleyebileceğiz.



simulation

- 276.484 benzersiz işlemden (transaction) oluşan ve 92.339 benzersiz öğe içeren 2.595.732 kayıt bulunmaktadır. Promosyon bilgilerini göz ardı ederek, insanların pazarlama çabalarından etkilenmek yerine ihtiyaç duydukları şeylere göre satın alım yapıldığı şeklinde varsayım ile simülasyon yapılmıştır.



Nichols, Elizabeth N. (2018). "Realistic Market Basket Data Simulation for Retail Marketing" Chancellor's Honors Program Projects.

Dataset

- Bu verisi kümesi şu değişkenleri içermektedir;

- **household_key:** Hanehalkı unique id'sidir.
- **Basket_ID:** İşlemlerin unique id'sidir.
- **Day:** işlemin yapıldığı gün
- **Product_ID:** Ürünlerin unique id'sidir.
- **Quantity:** Satın alınan miktar.
- **Sales_value:** Satış değeri
- **Store_id:** işlemin tamamlandığı mağaza kimliği
- **Retail_disc:** satın alınan ürünün perakende indirimi
- **Trans_time:** işlemin gerçekleştiği zaman
- **week_no:** işlemin gerçekleştiği hafta numarası
- **coupon_dısc:** kupon kullanım indirimini
- **coupon_match_dısc:** kupon kullanım eşleşme indirimi

transaction_data

household_key	BASKET_ID	DAY	PRODUC	QUANTITY	SALES_VALUE	STORE_ID	RETAIL_DISC	TRANS_TIME	WEEK_NO	COUPON_DISC	COUPON_MATCH_DISC
2375	26984851472	1	1004906	1	1.39	364	-0.6	1631	1	0	0
2375	26984851472	1	1033142	1	0.82	364	0	1631	1	0	0
2375	26984851472	1	1036325	1	0.99	364	-0.3	1631	1	0	0
2375	26984851472	1	1082185	1	1.21	364	0	1631	1	0	0
2375	26984851472	1	8160430	1	1.5	364	-0.39	1631	1	0	0
2375	26984851516	1	826249	2	1.98	364	-0.6	1642	1	0	0
2375	26984851516	1	1043142	1	1.57	364	-0.68	1642	1	0	0
2375	26984851516	1	1085983	1	2.99	364	-0.4	1642	1	0	0
2375	26984851516	1	1102651	1	1.89	364	0	1642	1	0	0
2375	26984851516	1	6423775	1	2	364	-0.79	1642	1	0	0
2375	26984851516	1	9487839	1	2	364	-0.79	1642	1	0	0
1364	26984896261	1	842930	1	2.19	31742	0	1520	1	0	0
1364	26984896261	1	897044	1	2.99	31742	-0.4	1520	1	0	0
1364	26984896261	1	920955	1	3.09	31742	0	1520	1	0	0
1364	26984896261	1	937406	1	2.5	31742	-0.99	1520	1	0	0
1364	26984896261	1	981760	1	0.6	31742	-0.79	1520	1	0	0
1130	26984905972	1	833715	2	0.34	31642	-0.32	1340	1	0	0
1130	26984905972	1	866950	2	0.34	31642	-0.32	1340	1	0	0

Simülatörler arasında sepet değerlerinin doğruluğunu karşılaştırmamıza olanak sağlamak için orijinal veri kümesinden bir Fiyatlar (Prices dataset) veri kümesi de oluşturulmuştur. Kuponlar ve promosyonlar ayrı kategorilerde listelendiğinden, her bir ürünün fiyatının haftadan haftaya çok fazla değişmeyeceği varsayılmıştır. Orijinal veri kümesindeki satış değeri (sales_value) değişkeni, satırlarda listelenen her bir öğenin "Fiyatlar" (Prices) sütununu oluşturmak için miktar değişkenine bölünmüştür. Sıfır miktarına bölmenin bir sonucu olarak NaN veya Infinity fiyatı olan herhangi bir öğe daha sonra 0,01 ABD Doları olarak değiştirildi

- Holdout: veri setinin eğitim ve test olmak üzere iki parçaya ayrıldığı yöntemdir. Test setinde kullanılan veri eğitim setinin dışındaki verilerden oluştuğu için bu yönteme holdout ismi verilmiştir.
- Holdout verisi olarak bir haftalık işlemleri seçmek için bir rasgele sayı üreteci kullanıldı, bu çalışma için, holdout verisi olarak kullanılmak üzere 12. hafta rasgele seçildi.
- İşlem formatına dönüştürülen verilerden, iki ayrı veri kümesi oluşturacak şekilde işlemlerin bir bölümü yapılmıştır,
- Biri 12. hafta işlemleri, diğeri 1-11 ve 13-102 hafta işlemleri olarak.
- İlk veri seti bizim holdout verisetimiz ve ikincisi eğitim verisidir. Bölme, orijinal verilerden Basket_ID'lerin listesinin transaction formatındaki verilerle aynı sırada düzenlenmesiyle yapılmıştır.

```
########################## #### Data Preparation #### ##########################


DH <- read.csv("/Users/merve.poslu/Downloads/transaction_data.csv")
head(DH)
## Create a transactions data object for Arules
DH_list <- split(DH[,"PRODUCT_ID"], # ItemID
DH[,"BASKET_ID"])  # TransID
DH_transactions <- as(DH_list,"transactions")
class(DH_transactions)

########################## #### Data Partitioning #### ##########################

# Randomly select a week for hold out
#WeekID <- sample(unique(DH$WEEK_NO),1)
HoldoutWeekID <- 12 # hold-out week of our choice
# List of all transaction IDs in the same order of DH_transactions
DH_transIDs <- names(DH_list)
# First, we will find the week-transID correspondence
TransWeek <- sqldf("SELECT DISTINCT WEEK_NO, BASKET_ID FROM DH")
# Then, we subset all transactions that happened in a particular week
HoldoutBaskets <- subset(TransWeek, WEEK_NO==HoldoutWeekID)
nrow(HoldoutBaskets) # number of hold-out baskets
nrow(TransWeek)-nrow(HoldoutBaskets) # number of training baskets
# ----- Splitting transactions into the training and Hold-out Sample Baskets -----
# İşlemleri eğitim ve hold-out örneklerine bölme---
#Holdout Baskets
SelectedTrans <- which(DH_transIDs %in% HoldoutBaskets$BASKET_ID)
DH_trans_holdout <- DH_transactions[SelectedTrans]
#Training Baskets
UnselectedTrans <- setdiff(1:length(DH_transactions), SelectedTrans)
DH_trans_train <- DH_transactions[UnselectedTrans]
```

Holdout verisi olarak bir haftalık işlemleri seçmek için rasgele 12. hafta seçildi.

Verilerin bir kopyası, işlemde satın alınan tüm Product_ID'lerin listesinin izlendiği benzersiz Basket_ID'lerin satırlarıyla yeniden biçimlendirildi.

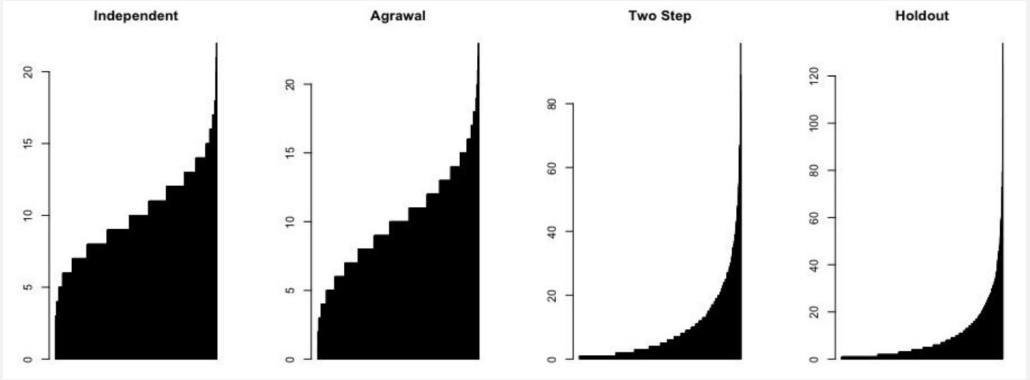
İşlem formatına dönüştürülen verilerden, iki ayrı veri kümesi oluşturacak şekilde işlemlerin bir bölümü yapılmıştır,

```
######## Data Simulation for Holdout Week ####

# number of unique items possible -- all possible items considered
numitems.all <- nrow(DH_transactions@itemInfo)
# number of transactions to simulate -- equals # trans in the holdout week
bSizes_actual <- size(DH_trans_holdout)
numtrans.ho <- length(bSizes_actual)
# ----- Independent Method -----
independentEXAMPLE <- random.transactions(numitems.all, numtrans.ho,method="independent
#Assuming default values
# ----- Agrawal Method -----
patterns <- random.patterns(numitems.all) #Assuming default values
agrawalEXAMPLE <- random.transactions(numitems.all, numtrans.ho, method="agrawal", patt
# ----- Two Step Method -----
# Step 1: simulate basket sizes
# Step 2: draw items for each basket
#Find out basket size distribution in training set
bSizes_train <- size(DH_trans_train)
itemFreq_train <- itemFrequency(DH_trans_train, type="absolute") #Frequencies of the it
itemFreq_train_prob <- itemFreq_train / sum(itemFreq_train) #Calculate probabilities of
#Sizes of each basket (to be simulated)
bSizes_2step <- sample(bSizes_train, numtrans.ho, replace=T)
simu_df <- NULL
for(i in 1:numtrans.ho){
  pick_items <- bSizes_2step[i]
  #Randomly pick these many items
  item_idx <- sample(1:length(itemFreq_train), pick_items,
                     prob=itemFreq_train_prob, replace=T)
  #Create a date frame of the randomly selected items and their corresponding transacti
  twostep <- data.frame(BASKET_ID=i,
  PRODUCT_ID=item_idx) #Combine the dataset
simu_df <- rbind(simu_df,twostep)
}
```

- Eğitim verileri kullanılarak, holdout verilerini denemek ve replike etmek için üç farklı simülatör çalıştırılmıştır. İlk ikisi, Independent ve Agrawal; diğeri ise Independent yöntemde fark edilen dezavantajları iyileştirmek için geliştirilmiş olan Two Step yöntemidir. KL Divergence iki olasılık dağılımını karşılaştırır.
- Bu üç ayrı simülasyonla, KL Divergence ölçüsü holdout için karşılaştırmalar yapılmıştır. KL Divergence, farkı ölçtüğü için veri setlerinin item frekansları, basket size yani sepet büyüklükleri ve sepet değerleri arasında karşılaştırmalar yapılmıştır. Bu aynı zamanda simülasyonlar arasındaki güçlü ve zayıf yönler hakkında daha kapsamlı bir görüşe sahip olmamızı sağlar.
- Simüle edilen veriler, R ARules paketindeki random.transactions fonksiyonu kullanılarak oluşturulmuştur. Hangi simülasyonun gerçek holdout verilerini kopyalamaya daha yakın olduğunu karşılaştırmak için hem Independent hem de Agrawal yöntemleri kullanıldı.

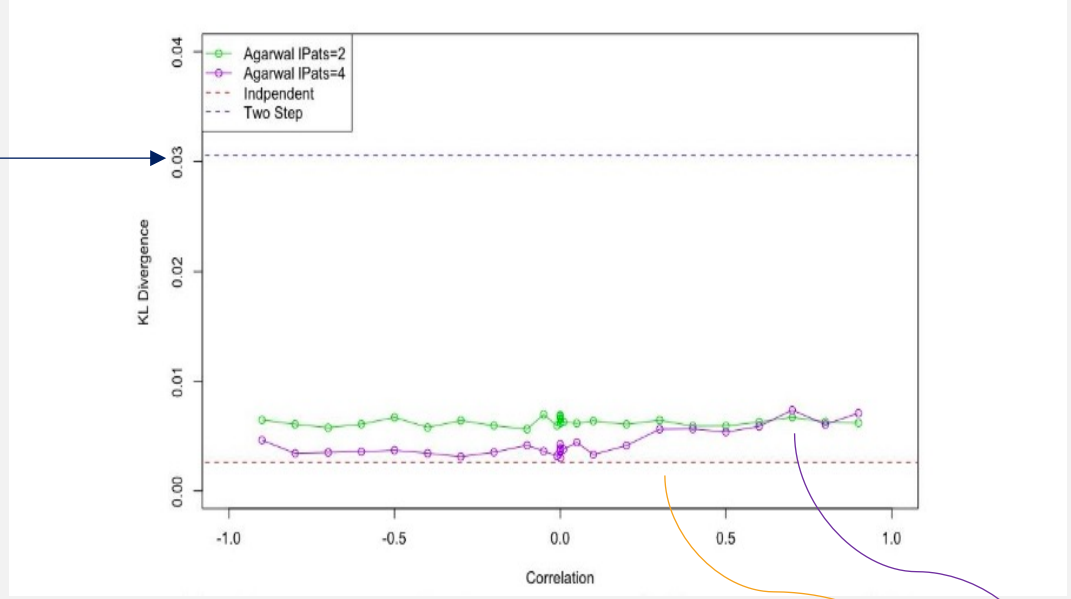
Öğe Frekans Dağılımları

```r
#### Evaluations ####
# ----- Item Frequency Comparisons -----
#Create vectors of item frequencies
itemFreq_actual <- itemFrequency(DH_trans_holdout, type="absolute")
itemFreq_simuI <- itemFrequency(independentEXAMPLE, type="absolute")
itemFreq_simuA <- itemFrequency(agrawalEXAMPLE, type="absolute")
itemFreq_simu2 <- itemFrequency(twostep_trans, type="absolute")

#Comparisons of the distributions of sorted item frequencies
par(mfrow=c(1,4))
barplot(sort(itemFreq_simuI), main="Independent")
barplot(sort(itemFreq_simuA), main="Agrawal")
barplot(sort(itemFreq_simu2), main="Two Step")
barplot(sort(itemFreq_actual), main="Holdout")
par(mfrow=c(1,1))

#Determine breaks for the KL Divergence computations
all_freqs <- c(itemFreq_actual,itemFreq_simuI,itemFreq_simuA,itemFreq_simu2)
breaks <- seq(min(itemFreq_actual),max(itemFreq_actual),length.out=50)

#Change the item frequecies into probabilities by the breaks
itemFreq_actual_distr <- hist(itemFreq_actual, breaks=breaks, plot = FALSE)$counts
itemFreq_simuI_distr <- hist(itemFreq_simuI, breaks=breaks, plot = FALSE)$counts
itemFreq_simuA_distr <- hist(itemFreq_simuA, breaks=breaks, plot = FALSE)$counts
itemFreq_simu2_distr <- hist(itemFreq_simu2, breaks=breaks, plot = FALSE)$counts
```



Not: Gösterim kolaylığı için R'dakinden daha büyük boyutlu, ayrıntılı gösterimdir.

Two Step yöntemin kurulmuş olan simülatörlere kıyasla Holdout yönteme daha fazla benzediğini potansiyel olarak daha iyi temsil edebilir, ancak tutarlılık için tüm karşılaştırmalar için KL Divergence kullanılır.



Independent veriler ile gerçek arasındaki KL Sapması 0.0026

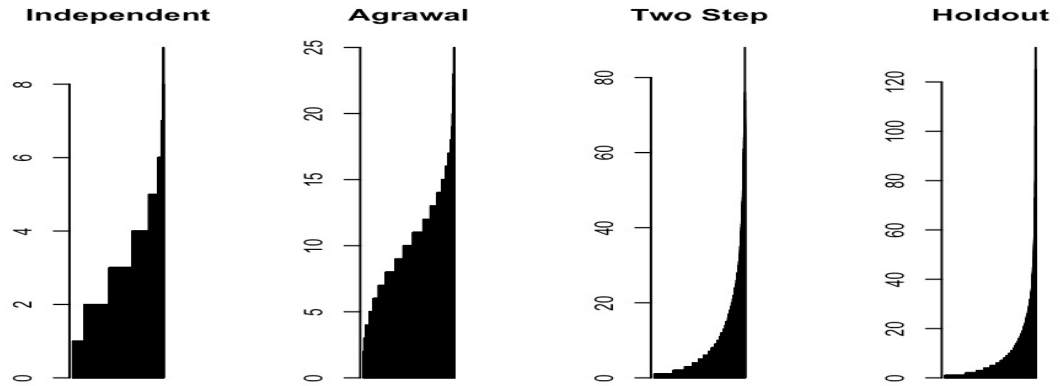
Agrawal verileri ile gerçek arasındaki KL sapması, yaklaşık 0,003 ile 0,007 arasında değişmiştir.

Two Step veri ile gerçek arasındaki KL Sapması 0.032 dür.

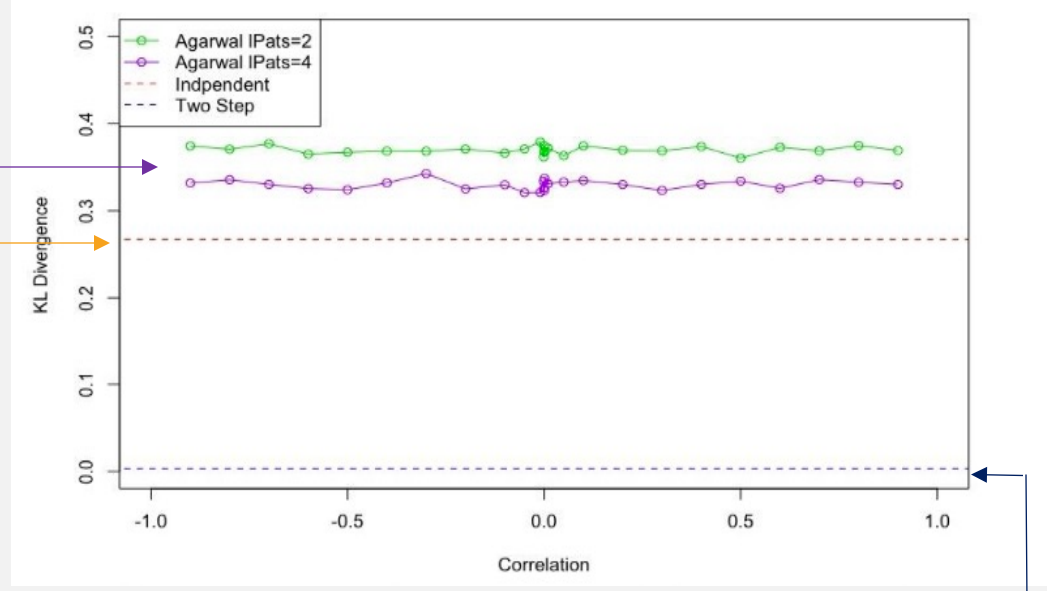
Holdout yöntemde görülen olasılıklarına göre öğeleri seçmek, öğe sıklıklarının şekli için en iyi eşleşmeyi sağlar. Ne yazık ki bu, KL sapmasına yansımaz, çünkü Two Step yöntem yalnızca karşılaştırma için sepetlere çektiği öğeleri içerirken, Independent ve Agrawal yöntemleri, sepetlerde olup olmadıklarına bakılmaksızın orijinal verilerdeki tüm öğeleri dikkate alır. Bu da, Two Step'de, frekans olasılığı sıfır olan tüm öğelerin simülasyondan hariç tutulduğu ve bu nedenle, sıfıra eşit olasılıkları içeren Holdout öğe frekanslarıyla karşılaştırıldığında kaçırıldığı anlamına gelir.

Sepet Büyüklüğü (Basket Size) Dağılımı

```
# ----- Basket Size Comparisons -----
#Create vectors of basket sizes
bSizes_indep <- size(independentEXAMPLE)
bSizes_agrawal <- size(agrawalEXAMPLE)
bSizes_2step <- size(twostep_trans)
bSizes_actual <- size(DH_trans_holdout)
#Comparisons of the distributions of sorted basket sizes
par(mfrow=c(1,4))
barplot(sort(bSizes_indep), main="Independent")
barplot(sort(bSizes_agrawal), main="Agrawal")
barplot(sort(bSizes_2step), main="Two Step")
barplot(sort(bSizes_actual), main="Holdout")
par(mfrow=c(1,1))
```



Two Step simülasyonun, sepet büyüklüklerinin uygun dağılımını doğru bir şekilde simüle etmede Independent ve Agrawal yöntemlerinden önemli ölçüde daha iyi olduğu görülebilir. Holdout, 120'den fazla öğe içeren en büyük sepetlerle üstel dağılıma benzer bir dağılım izler. Two Step, aynı dağılımı izler ve yalnızca en büyük sepetleri kaçırırken, diğer iki yöntem hem dağılımı hem de orta ve büyük sepet boyutlarını kaçırır.



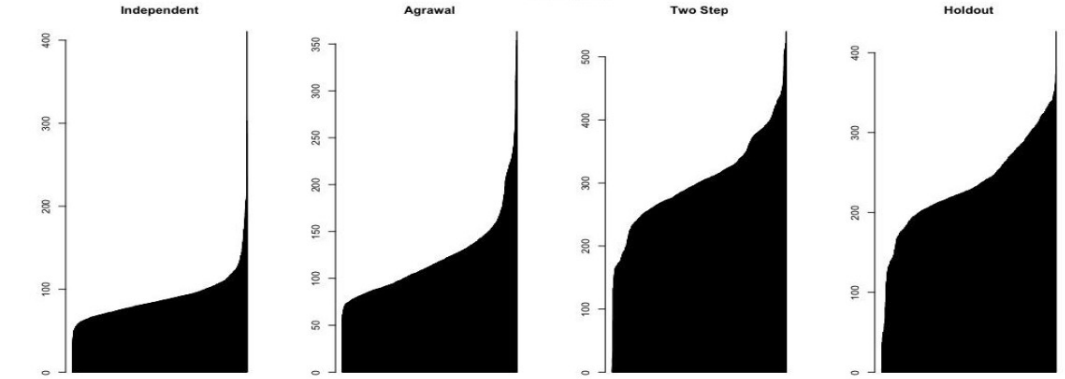
- Independent veriler ile gerçek arasındaki KL Farklılığı 0.27 dir.
- Agrawal verileri ile gerçek arasındaki KL Sapması yaklaşık 0,3 ile 0,4 arasında değişiyordu ve 4'lük model ortalaması 2'lik model ortalamasından daha iyi performans gösteriyordu.
- Two Step veri ile gerçek arasındaki KL Ayrımı 0,003 dür.

Buradaki KL Divergence, 0'a son derece yakın bir değerle Two Step'in ne kadar iyi performans gösterdiğini algılar.
Poisson dağılımını varsaydıkları için daha büyük sepetleri simüle edemeyen Independent ve Agrawal, Two Step'in değerinden 0.25'in üzerinde KL Divergence değerlerine sahiptir.

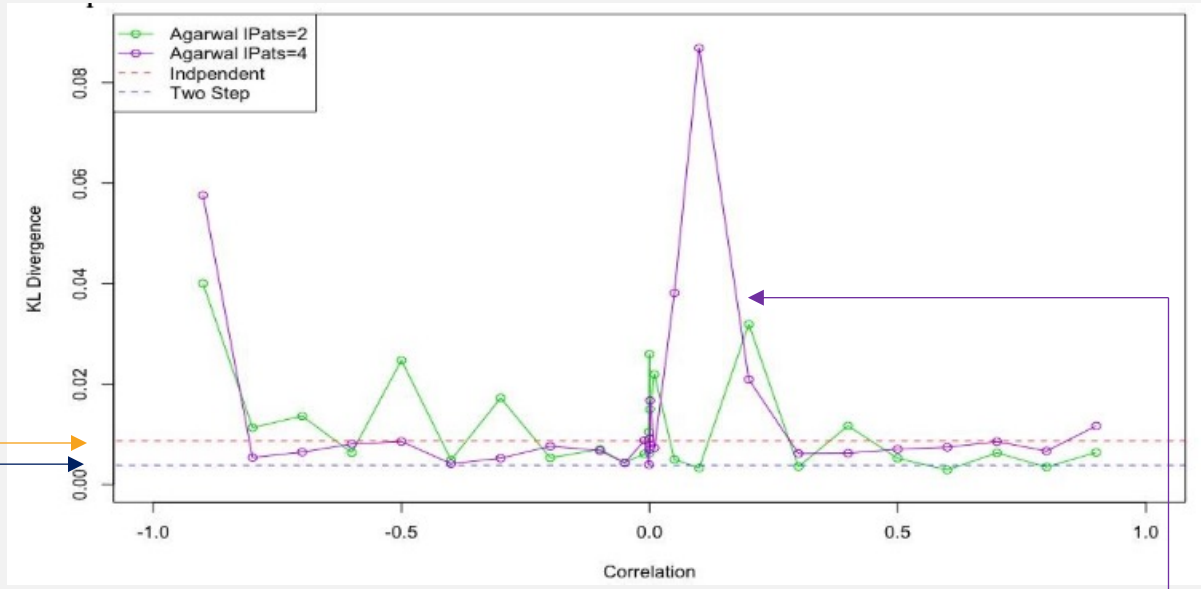
Sepet Değer Dağılımı

```
# ----- Basket Value Comparisons -----
#Create Prices dataset for items
DH$Price <- DH$SALES_VALUE/DH$QUANTITY #Create price for each item individually
DH$Price[is.nan(DH$Price)] <- 0.01 #Change any value of NaN to a penny
DH$Price[is.infinite(DH$Price)] <- 0.01 #Change any value of infinity to a penny
Prices <- aggregate(Price~PRODUCT_ID, data=DH,mean) #Create general prices by the mean

#Convert Agrawal and Independent items IDs to UPCs
item_labels <- DH_transactions@itemInfo
iLabels <- itemLabels(DH_transactions)
list <- LIST(DH_trans_holdout, decode = FALSE)
baskets <- list
list <- decode(list, itemLabels = iLabels)
baskets <- decode(baskets, itemLabels = iLabels)
HO_baskets <- as(baskets,"matrix")

list2 <- LIST(agrawalEXAMPLE, decode = FALSE)
baskets2 <- list2
list2 <- decode(list2, itemLabels = iLabels)
baskets2 <- decode(baskets2, itemLabels = iLabels)
AGR_baskets <- as(baskets2,"matrix")

list3 <- LIST(independentEXAMPLE, decode = FALSE)
baskets3 <- list3
list3 <- decode(list3, itemLabels = iLabels)
baskets3 <- decode(baskets3, itemLabels = iLabels)
IND_baskets <- as(baskets3,"matrix")

list4 <- LIST(twostep_trans, decode = FALSE)
baskets4 <- list4
list4 <- decode(list4, itemLabels = iLabels)
baskets4 <- decode(baskets4, itemLabels = iLabels)
TwoStep_baskets <- as(baskets4,"matrix")
```

Matristeki UPC'leri kendi fiyatlarıyla değiştirmek ve ardından sepetteki tüm fiyatları toplamak için bir fonksiyon oluşturuldu. Bu fiyat toplamları, KL Divergence ile karşılaştırmak için bir vektör formunda kaydedilen sepetlerin değerlerini temsil eder.



Ayrıca, dağılım grafiklerinin karşılaştırılmasıyla, Two Step'in, üç simülatörden holdout dağılımıyla en iyi şekilde eşleştiği görülebilir.



- Independent veriler ile gerçek arasındaki KL Farkı 0,026 dır.
- Agrawal verileri ile gerçek arasındaki KL Sapması yaklaşık 0,00 ile 0,09 arasında değişiyordu, model ortalaması 4 ve model ortalaması 2 korelasyona bağlı olarak birbirinden daha iyi ve daha kötü performans gösteriyor gibi görünüyor.
- Two Step veriler ile gerçek veriler arasındaki KL Sapması 0.00412 dir.
- Buradaki KL Divergence değerlerinin dağılımı, önceki iki karşılaştırma alanından Agrawal yöntemi için çok daha çeşitlidir ve bunun neden meydana geldiğine dair sağlam bir açıklama yoktur. Sebeplerden biri, aralık o kadar küçük olabilir ki, korelasyonlar arasındaki küçük değişiklikler grafik tarafından daha belirgin bir şekilde algılanıyor olabilir. Daha gerçekçi bir açıklama ise, her sepetteki öğeler sepet değerini belirlediğinden, yüksek veya düşük değerli herhangi bir öğe toplam sepet fiyatını gerçekten çarpıtabileceğinden, burada daha fazla varyasyon olması muhtemeldir.
- Two Step hala Agrawal ve Independent'tan daha iyi performans gösteriyor

- **SONUÇ**
- Üç karşılaştırma noktasının dağılımlarına bakıldığında, üç değerlendirme alanında Two Step simülasyonun gerçek verilerle en doğru şekilde eşleştiği görülmektedir.
- Sıralanan verilerle aynı şekilde en yüksekleri ve en düşükleri alır ve her bir grafik kümesinden görsel olarak en çok benzeyen dağılıma benzer. Bu görsel sonuç, karşılaştırılan değişkenden bağımsız olarak Two Step için sürekli olarak düşük KL Divergence değerlerinde sayısal olarak yansır.
- Sonuç olarak, Agrawal ve Independent, büyük sepetleri simüle edemez çünkü bunlar, kodunda yazılan set dağılımlarıyla sınırlıdır. Two Step yöntemin yararı, dağılımın, çoğu durumda holdout verilerinin daha geniş bir temsili olan eğitim verileri tarafından oluşturulmasıdır. Bu, varolan yöntemlerin eksikliğine yönelik bir düzeltmedir. Bu düzeltme ile birlikte Two Step'te de bir dezavantaj vardır; varolan yöntemleri temsil ederken olası tüm öğeleri temsil etmez.
- Çok sayıda benzersiz öğe içeren perakende verileriyle, simülasyonda dikkate alınmaya değer herhangi bir güçlü öğe çifti korelasyonuna sahip olmak için işlemler içinde çok fazla olası öğe kombinasyonu vardır. Sonuç olarak, Agrawal yöntemi çok karmaşıktır.
- Öğeler arasında herhangi bir korelasyon olmayan Independent yöntem, gerçek perakende dünyasında nadiren görülen tekdüzeliği varsaydığı için çok basittir.
- Two Step yöntemin üç karşılaştırma değişkeni arasında nasıl iyi performans gösterdiğini görmek, perakende verilerini simüle etmenin en iyi yolu olduğunu doğrular.