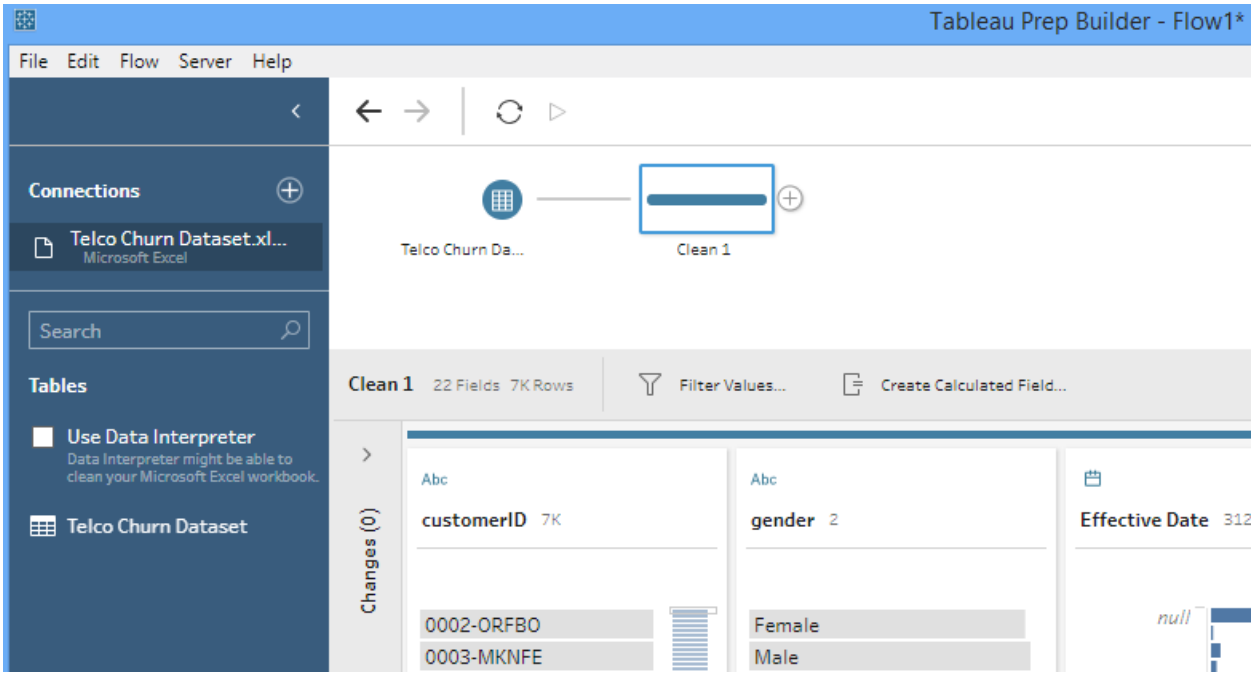
Tableau Prep Builder – Exercitii de curatare a datelor

Tableau Prep Builder oferă funcţionalităţi în pregătirea datelor pentru analiză, ajutând la creşterea performanţei şi a optimizării extragerii datelor.

Vom utiliza pentru exemplificare setul de date *Telco Churn Dataset.xlsx* .

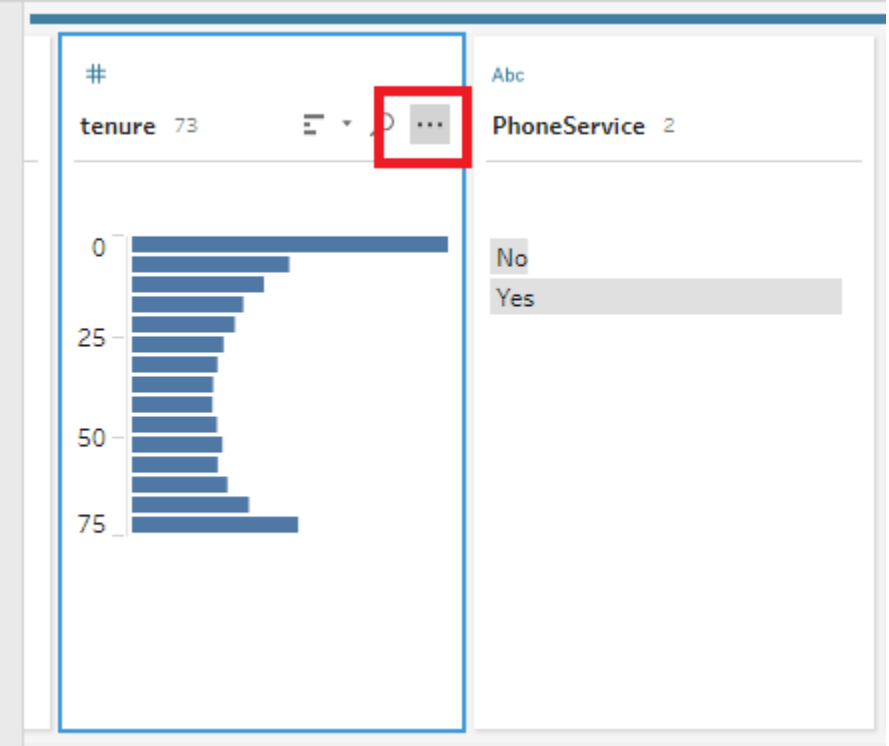
Setul de date folosit se referă la una din problemele comune în domeniul telecomunicaţiilor, respectiv migrarea(portarea) clienţilor. Conform mai multor studii, a fost demonstrat că un client nou costă mult mai mult decât păstrarea unuia vechi. Aşadar companiile sunt interesate în dezvoltarea unor modele de identificare a clienţilor cu probabilitate ridicată să renunţe la servicii.

Înainte de a realiza operaţii în Tableau Prep, este recomandat să vizualizăm datele, respectiv structura lor şi să înţelegem semnificaţia variabilelor. Dupa ce conectam setul de date vom adauga un pas “Clean” pentru a avea posiblitatea de vizualizare a structurii datelor.



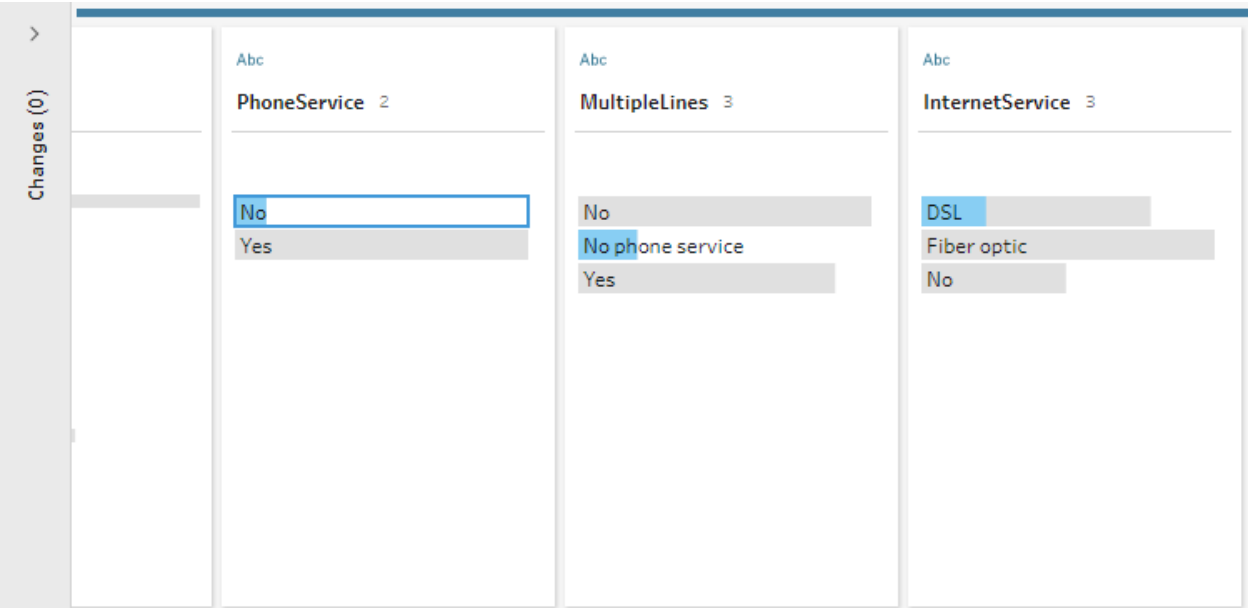
Observati modul de reprezentare a variabilelor discrete din setul de date (Ex. PhoneService), în toate valorile sunt afişate(No/Yes), iar frecvenţa lor este reprezentată de mărimea barelor orizonatele gri.

Variabilele continue (Ex. tenure) sunt reprezentate sub forma unor grafice de tip histogramă, unde valorile sunt grupate pe intervale, iar distribuţia lor este reprezentată prin bare orizontale albastre.



Acest format de vizualizare a variabilelor continue poate fi modificat prin selectarea butonului More Options si apoi a optiunii Detail, caz în care vor fi afişate toate valorile, similar cu variabilele discrete.

Afişarea din acest panel ne poate ajuta să observăm imediat anumite informaţii din setul de date, prin selectarea unor categorii. De exemplu selectând valoarea No pentru variabila PhoneService, putem observa că toate aceste personae au InternetService de tip DSL.



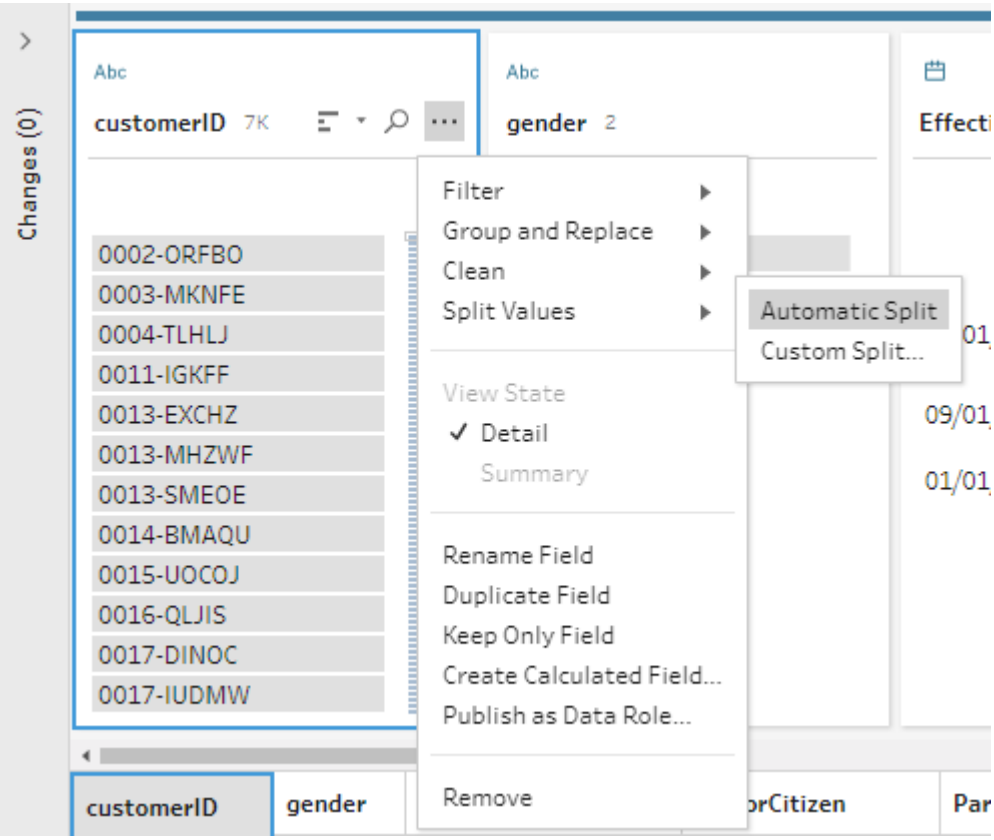
Funcţii

Avem posibilitatea de a realiza operaţii de curăţare a datelor în urma vizualizării lor, incluzând

Fitrare, Splitare, Redenumirea câmpurilor, Eliminarea câmpurilor, Schimbarea tipului datelor etc.

1) Splitare

În general în seturile date există situaţii în care un câmp conţine mai multe informaţii (ex. Data calendaristică, CNP). Aceste informaţii pot fi extrase sub forma unor câmpuri noi atunci când considerăm ca ne sunt utile. În acest set de date, câmpul customerID este format din două părţi, separate prin caracterul “-“. Prima parte reprezintă un cod intern, iar a doua codul de client. Putem realiza splitarea în mod automat prin intermediu funcţiei Split Values.



Vizualizaţi rezultatul splitării şi diferenţa dintre cele două câmpuri nou create.

2) Redenumirea câmpurilor

Uneori câmpurile trebuie redenumite pentru a avea un înţeles mai bun pentru utilizatori. Modificaţi denumirile celor două câmpuri create mai sus prin splitare, folosind opţiunea Rename Field.

3) Eliminarea câmpurilor

Pentru a elimina anumite câmpuri ne putem folos de opţiunea Rmove. Eliminarea câmpurilor este utilă pentru a crea un set de date mai suplu, cu mai puţină memorie ocupată inutil.

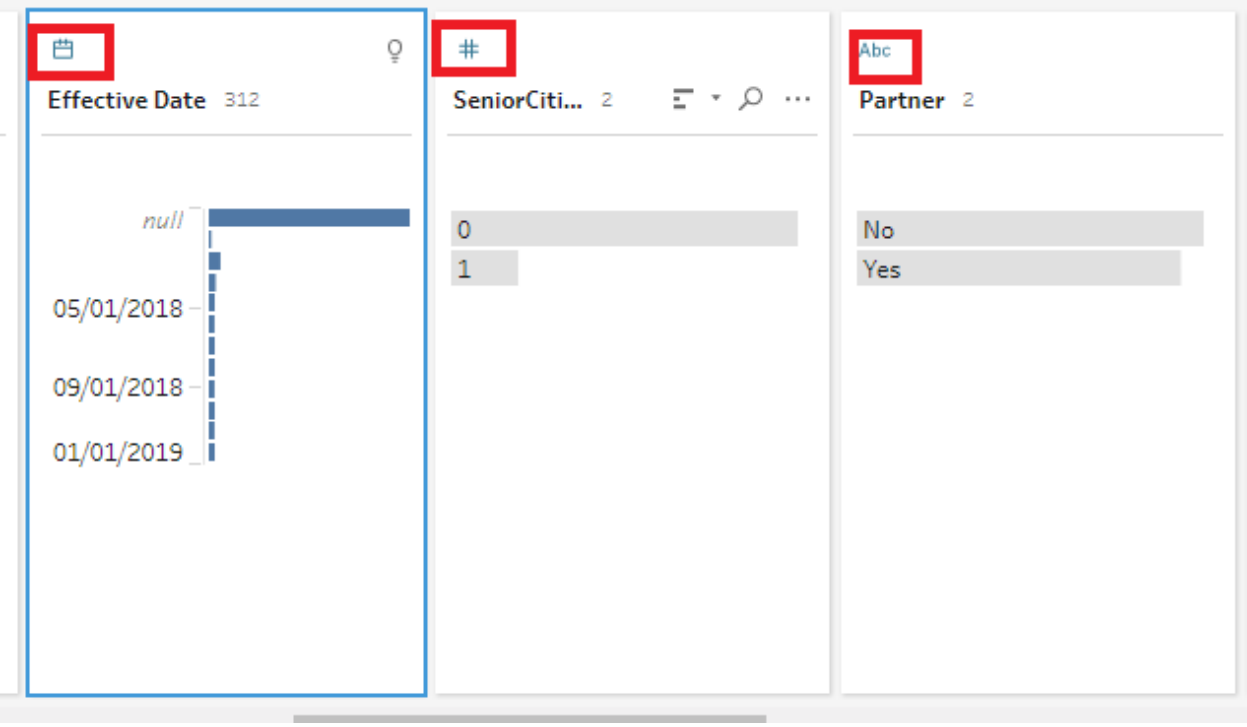
Eliminaţi câmpul customerID din exemplul de mai sus.

4) Schimbarea tipurilor datelor

Tableau Prep are definite următoarele tipuri de date:

- Number(decimal)
- Number(whole)
- Date and time
- Date
- String

La conectarea unei surse de date, tipurile sunt alocate în mod automat în funcţie de conţinutul fiecărui câmp. In situatia in care alocarea automata a tipului nu este satisfacatoare, avem posibilitatea de a modifica in mod manual tipul. Aceasta operatiune se face apasand pe pictograma din stanga sus a campului.



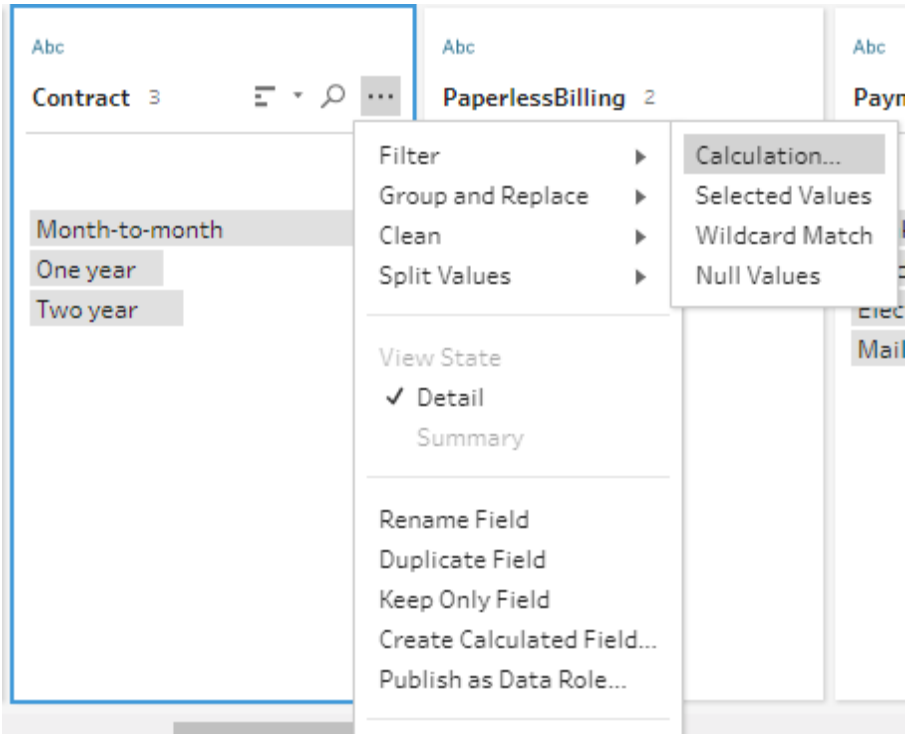
5) Filtrarea datelor

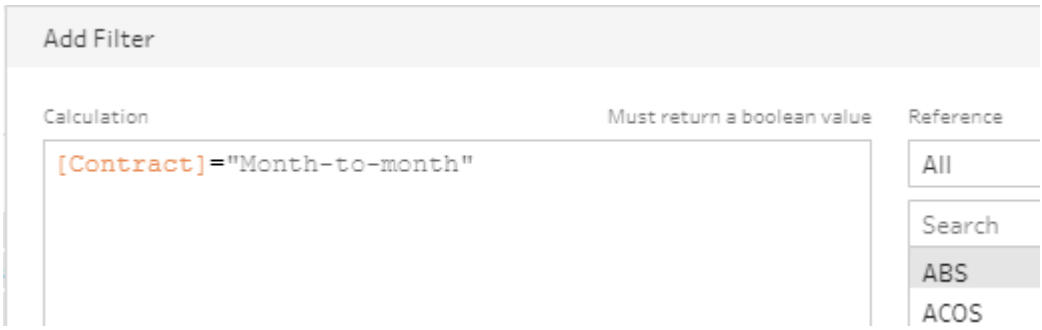
Filtrarea datelor ajută utilizatorul în a elimina informaţiile care nu îi sunt necesare şi a le păstra numai pe cele relevante.

În exemplul folosit în acest laborator presupunem că ne dorim să identificăm în baza de date numai un grup de clienţi care respectă anumite criterii. Aceste criteria sunt:

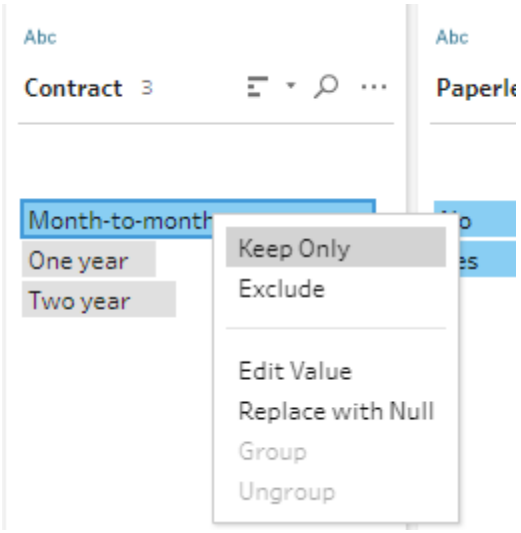
- Sa fie un client cu contract semnat in ultimele doua luni (din setul de date)
- Sa aiba un contract de tip month to month
- Sa genereze venituri mai mici de 80$ pe luna

Filtrele pot fi adaugate pentru fiecare câmp sub formă de calculation. Spre exemplu, tipul contractului month to month.



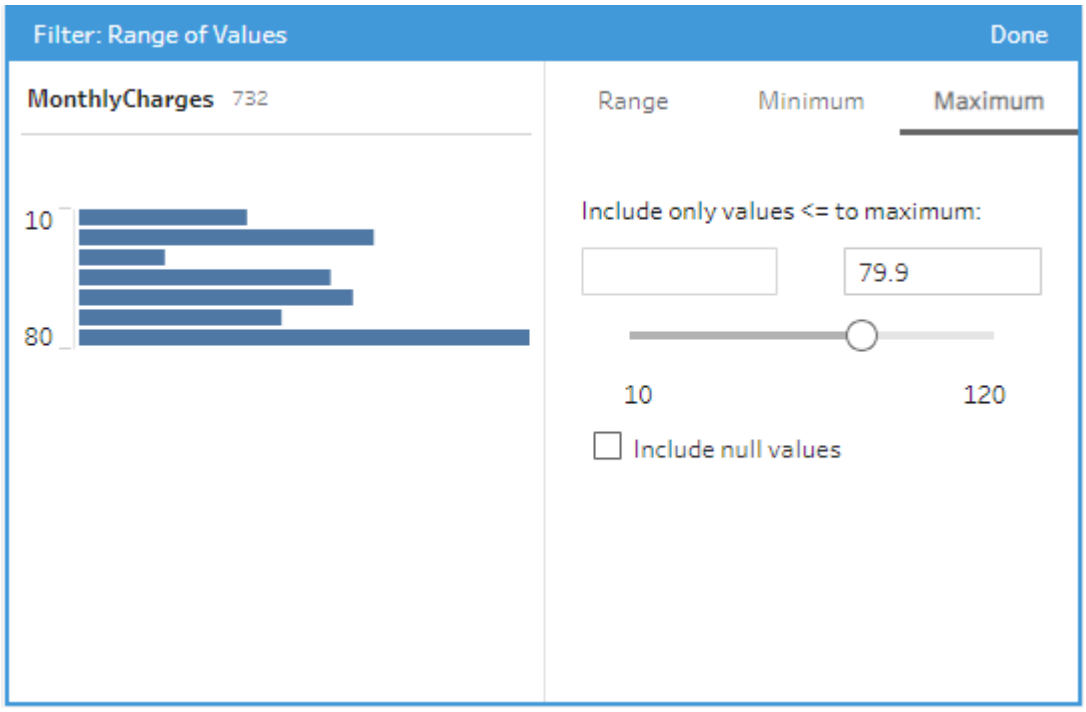


Un mod similar de a realiza acelaşi lucru este prin a selecta valoarea dorită pentru câmpul respectiv şi apoi a selecta Keep only.



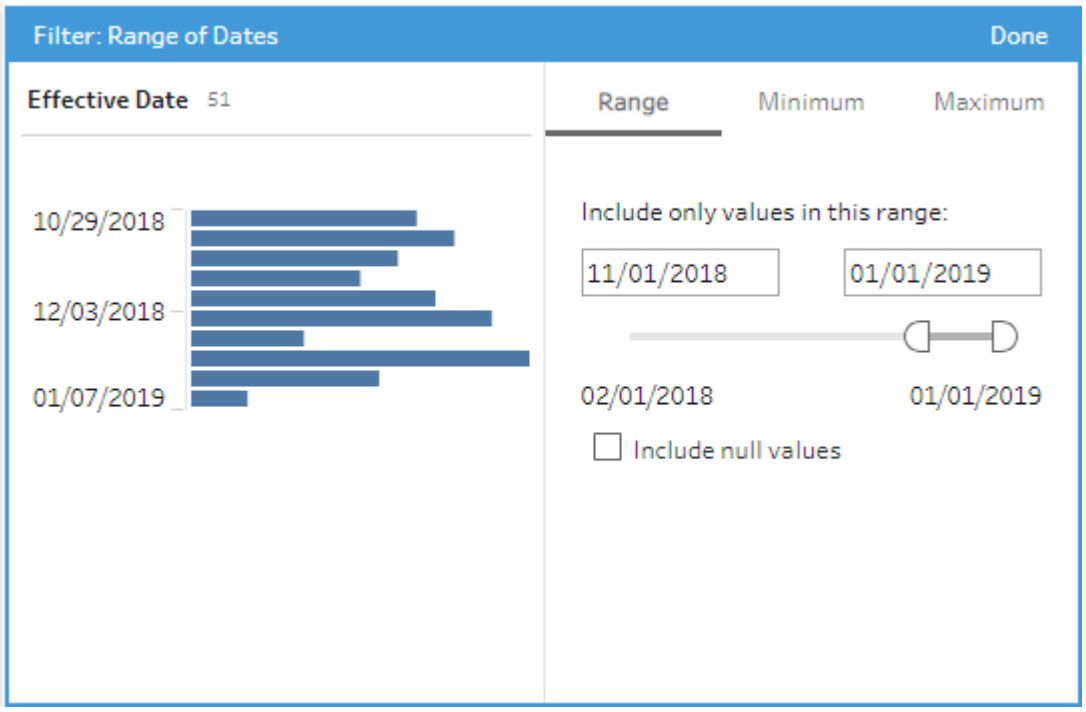
Realizati filtrarea pentru conditia ca MonthlyCharges sa fie mai mici de 80$ folosind Calculation, ca in exemplul de mai sus.

O alta modalitate pentru aceasta filtrare este aceea de a utiliza Range of Values si a selecta in mod vizual valorile dorite.



Valorile nule din setul de date pot fi sau nu incluse în rezultate, în funcţie de decizia utilizatorului. Spre exemplu, ar trebui să ne întrebăm dacă în această situaţie valorile nule reprezintă clienţi care nu aduc venituri sau clienţi pentru care nu se cunoaşte venitul.

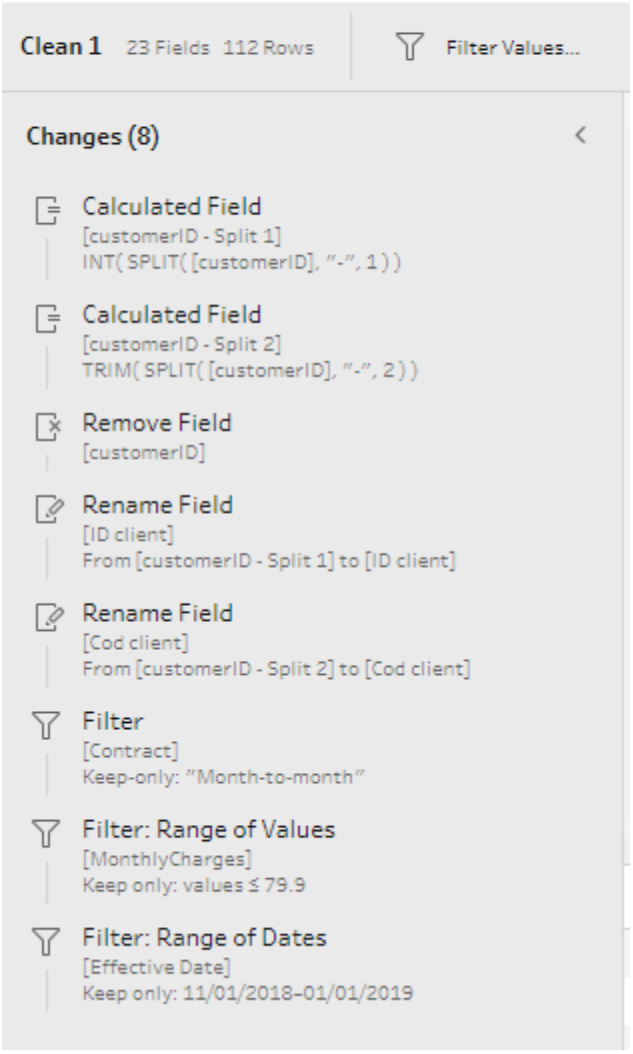
Pentru a filtra in funcţie de Data contractului, respectiv ultimele două luni din setul de date, vom folosi filtrarea pentru câmpul Effective Date si vom selecta Range of Dates.



Vizualizarea modificărilor realizate

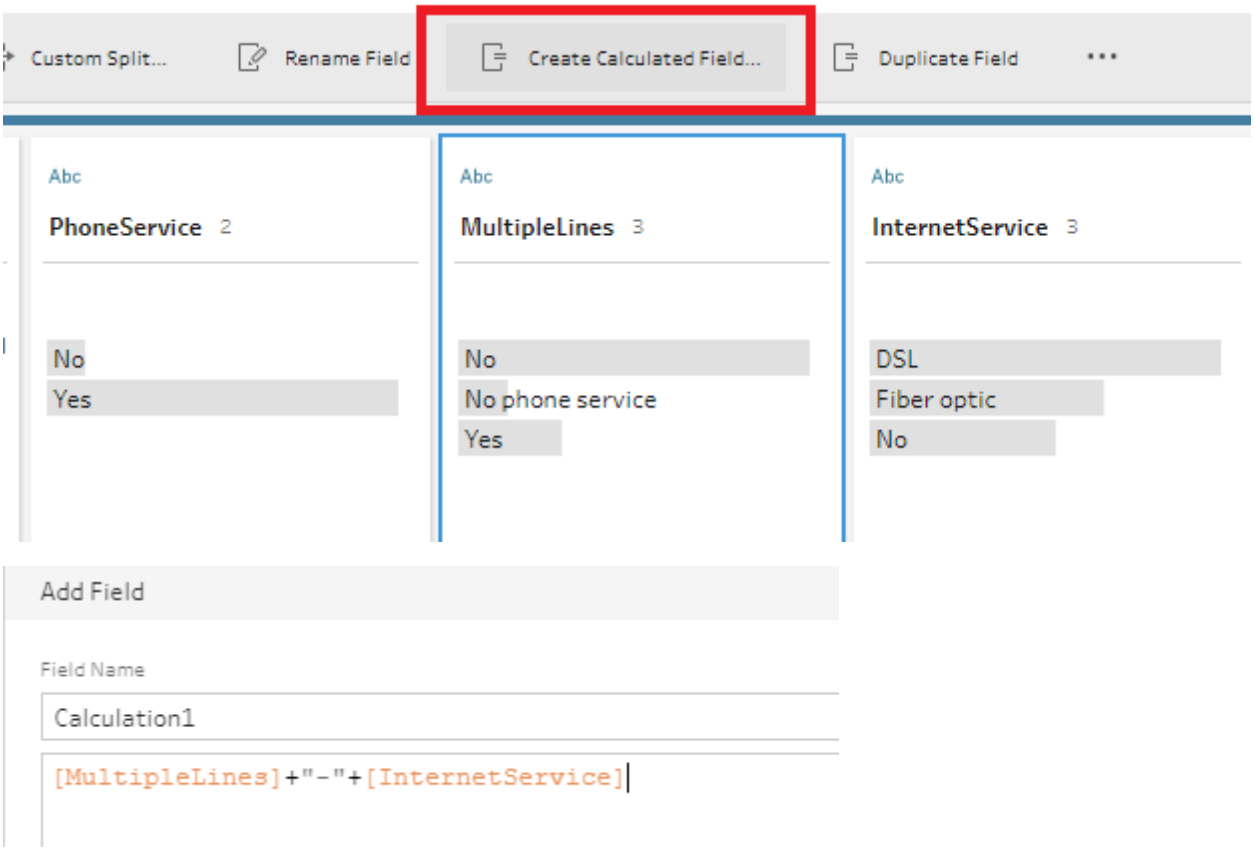
Vizualizaţi modificările realizate în panelul Changes. Daca dorim sa anulam unele dintre modificarile facute putem sa le facem din aceasta interfata.

Atentionare: Ordinea modificarilor facute este aceea indicata de sus in jos, asadar rezultatele pot fi diferite in functie de aceasta.



Introducerea unor campuri calculate

Este util uneori sa introducem noi câmpuri în setul de date, calculate pe baza altora existente. Să presupunem că dorim să introducem un nou câmp care să combine câmpurile MultipleLines şi InternetService, astfel încât să vedem în acelaşi câmp dacă clienţii utilizează cele două tipuri de servicii. Pentru aceasta vom utiliza opţiunea Create Calculated Field si vom realiza o operaţiune de concatenare a valorilor celor două câmpuri.



Observaţi modul în care noul câmp creat a concatenate valorile celorlalte două.

După parcurgerea acestor operaţii adăugaţi un nou pas de Output pentru a putea salva datele în formatul dorit.

Observaţi deasupra pasului Clean anumite pictograme care reprezintă modificările realizate asupra setului de date în paşii descrişi mai sus. Pentru a vizualiza oricare dintre modificări putem să selectăm pictograma dorită şi vor fi afişate.

Exerciţii

Folosind setul de date Telco Churn Dataset si prelucrarile realizate in cadrul laboratorului, parcurgeţi următoarele cerinţe:

a) Anulaţi filtrarea realizată pentru a selecta clientii cu contract Month to month
b) Extrageţi clienţii care utilizează Fibra Optică si nu folosesc Paperless Billing
c) Adaugati un pas de tip Aggregate pentru a observa media sumelor platite de clientii care s-au portat si de cei rămaşi (câmpurile Monthly Charges şi Churn)
d) Copiaţi şi Adaugati acelasi pas Aggregate în paralel cu pasul Clean de la punctual b si comparati valorile pentru intreg setul de date cu valorile de la punctual c.
e) Adaugati in cadrul pasului de la punctul d si campul Total Charges, pentru care sa afisati media. Ce puteti spune despre proporţiile celor doua categorii(Yes şi No)?
f) Folosind un alt pas Aggregate calculaţi media lunară plătită de bărbaţi si femei.
g) Exportaţi fiecare din rezultatele de mai sus ale pasilor Aggregate in format Tableau Data Extract (.hyper) si importaţi-le în Tableau Desktop (4 fisiere).
h) Realizaţi câte un grafic, la alegere, folosind datele importate.