

ArcGIS Insights Platfromunda Data Engineering Uygulaması

Veri türünü değiştirmek veya verilerimizi filtrelemek için Insights'ta çeşitli araçlar vardır. Verilerimizi önceden işlememizi sağlamak, analizimizi kolaylaştıracaktır. Insights desktop 2022.2'de, preview'da olan Veri Mühendisliği (Data Engineering) adlı yeni bir bölüm göreceğiz. Veri mühendisliği önizlemesi bize tam işlevli bir yetenek silsilesi sunar. Ancak bu durum şu anda yalnızca desktop sürümünde mevcuttur.

Hızlıca Uygulama safhasına geçelim.

Uygulama'da Veri Mühendisliği açısından gezineceğimiz ana temalar;

- Data Import
- Data Exploring
- Data Cleaning
- Statistic
- Write Data

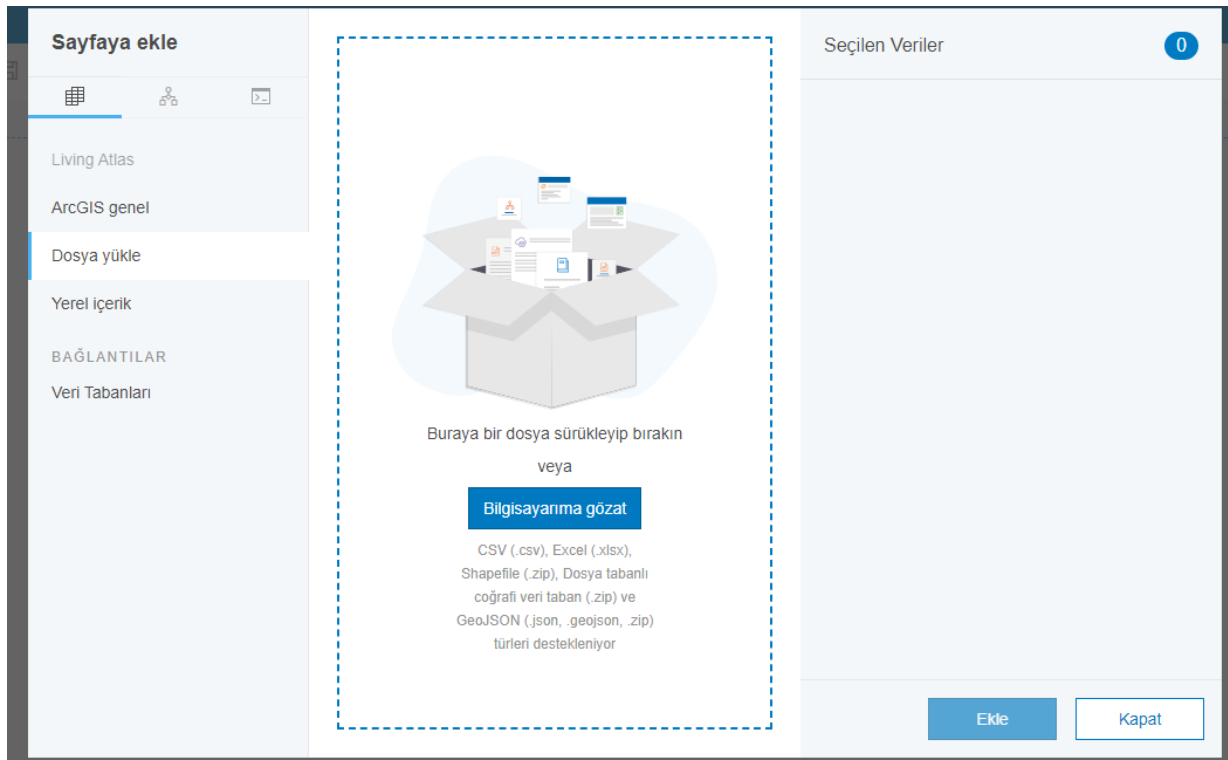
Kullanacağımız Veri Seti;

- Kaggle platfromundan çekilen Netflix veri seti

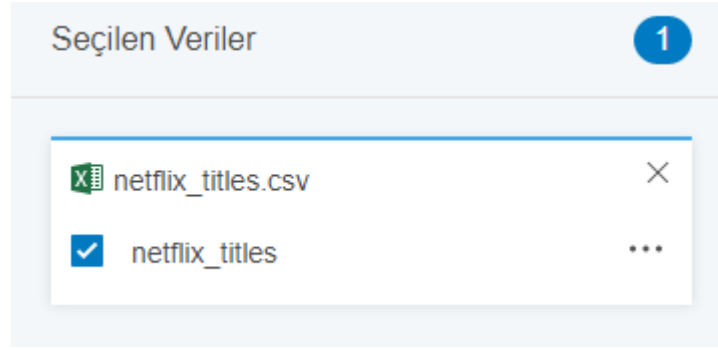
Kaggle platfromundan çekilen Netflix veri setinde ArcGIS Insights Data Engineer çalışması aşamaları;

1.İlk aşamada Data Engineer kategorisinden çalışma kitabı yarattık.

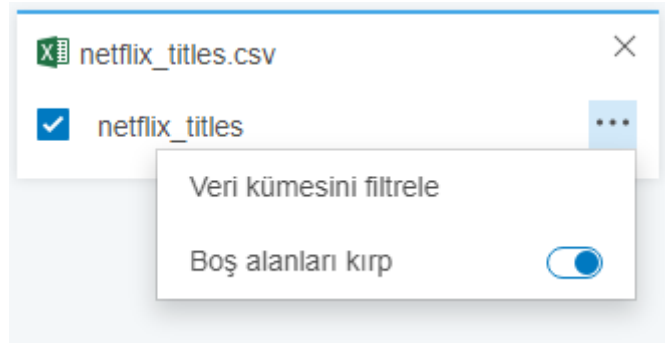
2.Yaratılan çalışma kitabına veri eklemek için sayfaya ekle butonundan işleyeceğimiz veriyi ekleriz.



3. Veri uygun lokasyondan seçilir ve aşağıdaki ayrıntı karşımıza çıkar. Bu işlem Import Data aşamasına bağlıdır.



Veri başlığı yanındaki noktali simgeye tıkladığımızda karşımıza “Veri kümesini filtrele” ve “Boş alanları kırp” araçları çıkacaktır.



Bu iki araç veri keşfi aşamasında bize çok yardımcı olacaktır.

Veri Kümesini Filtrele Aracı

-Bu araç ilk aşamada verideki sütunları (değişkenleri) filtrelememizi sağlar.

Tikli kısımların tiklerini kaldırdığımızda R ya da pythondaki “select “ fonksiyonuna muadil bir durum oluşur.

Seçilen sütunlar	12	Ön izleme
Seç		<input checked="" type="checkbox"/> show_id <input checked="" type="checkbox"/> type <input checked="" type="checkbox"/> title <input checked="" type="checkbox"/> director <input checked="" type="checkbox"/> cast
<input type="text" value="Arama"/>		
Filtre		
<input checked="" type="checkbox"/> netflix_titles		
<input checked="" type="checkbox"/> show_id		
<input checked="" type="checkbox"/> type		
<input checked="" type="checkbox"/> title		
<input checked="" type="checkbox"/> director		
<input checked="" type="checkbox"/> cast		
<input checked="" type="checkbox"/> country		
<input checked="" type="checkbox"/> date_added		
<input checked="" type="checkbox"/> release_year		
<input checked="" type="checkbox"/> rating		
<input checked="" type="checkbox"/> duration		
<input checked="" type="checkbox"/> listed_in		
<input checked="" type="checkbox"/> description		
<input type="button" value="İptal"/>	<input type="button" value="Bitir"/>	
		Toplam Kayıt Sayısı: 250

	show_id	type	title	director	cast
1	s1	TV Show	3%		João Miguel, Bianca
2	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor
3	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella C
4	s4	Movie	9	Shane Acker	Elijah Wood, John C
5	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin
6	s6	TV Show	46	Serdar Akar	Erdal Beşikçioğlu, Y
7	s7	Movie	122	Yasir Al Yasiri	Amina Khalil, Ahme
8	s8	Movie	187	Kevin Reynolds	Samuel L. Jackson,
9	s9	Movie	706	Shravan Kumar	Divya Dutta, Atul K
10	s10	Movie	1920	Vikram Bhatt	Rajneesh Duggal, A
11	s11	Movie	1922	Zak Hilditch	Thomas Jane, Moh
12	s12	TV Show	1983		Robert Więckiewicz
13	s13	TV Show	1994	Diego Enrique Osorno	
14	s14	Movie	2,215	Nottapon Boonprakob	Artiwara Kongmalai
15	s15	Movie	3022	John Suits	Omar Epps, Kate W
16	s16	Movie	Oct-01	Kunle Afolayan	Sadiq Daba, David
17	s17	TV Show	Feb-09		Shahd El Yaseen, S
18	s18	Movie	22-Jul	Paul Greengrass	Anders Danielsen L
19	s19	Movie	15-Aug	Swapnaneel Jayakar	Rahul Pethe, Mrun
20	s20	Movie	'89		Lee Dixon, Ian Wri
21	s21	Movie	Kuch Bheege Alfaaz	Onir	Geetanjali Thapa, Z

-Gelişmiş butonu ise veride filter işleminin yapılmasını sağlar yani bu ne demek oluyor?
R ve Python gibi programlama dillerinde veri keşfi aşamasındaki filter fonksiyonuna karşılık geliyor yani,gözlem değerlerini filtreleyerek tabloyu daraltmak anlamına geliyor.

Boş Alanları Kırp Aracı;

-String değerlerdeki boş alanları kırpar yani trim fonksiyonunun görevini gerçekleştirir.Bu özellik varsayılan olarak aktif gelir.

Veriyi çalışma alanımıza kaydettikten sonra Model ve Tablo görünümü şeklinde aşağıdaki gibi görmüş oluruz.

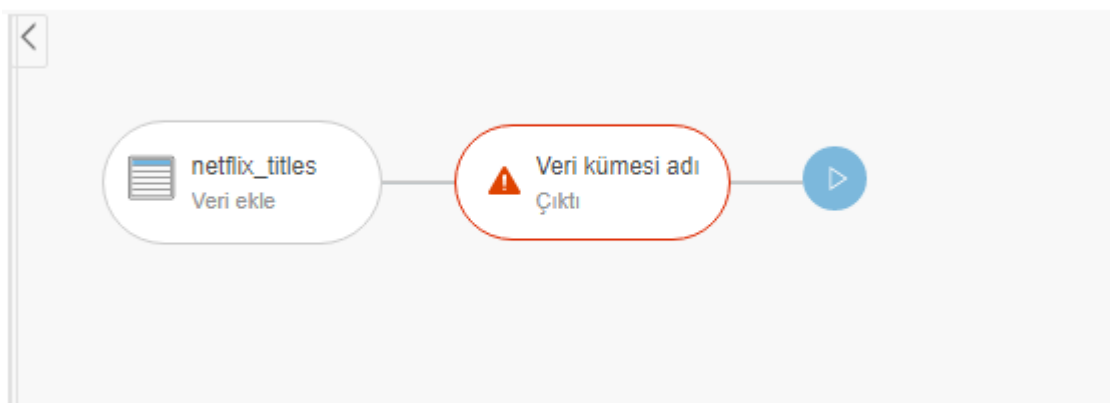
Bu ekran bizim veri mühendisliği çalışmalarımızın genel görünümüdür.Oluşturulan yeni verileri görselleştirmeye hazırlama aşaması olarak kullanırız.

The screenshot shows the Netflix Data Science interface. On the left, there's a sidebar with a search bar and a list of data clusters. The main area displays a workflow diagram with a data source 'netflix_titles' and a filter step 'Veri kümesi adı Çıktı'. Below the diagram is a table of data.

	show_id	type	title	director	cast	country
1	s1	TV Show	3%		João Miguel, Bianca C...	Brazil
2	s2	Movie	7:19	Jorge Michel Grau	Demian Bichir, Héctor ...	Mexico
3	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chu...	Singapore
4	s4	Movie	9	Shane Acker	Elijah Wood, John C. R...	United States
5	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Sp...	United States
6	s6	TV Show	46	Serdar Akar	Erdal Beşikçioğlu, Yase...	Turkey
7	s7	Movie	122	Yasir Al Yasiri	Amina Khalil, Ahmed D...	Egypt
8	s8	Movie	187	Kevin Reynolds	Samuel L. Jackson, Jo...	United States
9	s9	Movie	706	Shravan Kumar	Divya Dutta, Atul Kulka...	India
10	s10	Movie	1920	Vikram Bhatt	Rajneesh Duggal, Ada...	India

Toplam Kayıt Sayısı: 7.787

Model görünümü;



Tablo görünümü;

netflix_titles X						
	show_id	type	title	director	cast	country
1	s1	TV Show	3%		João Miguel, Bianca C...	Brazil
2	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor ...	Mexico
3	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chu...	Singapore
4	s4	Movie	9	Shane Acker	Elijah Wood, John C. R...	United States
5	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Sp...	United States
6	s6	TV Show	46	Serdar Akar	Erdal Beşikçioğlu, Yase...	Turkey
7	s7	Movie	122	Yasir Al Yasiri	Amina Khalil, Ahmed D...	Egypt
8	s8	Movie	187	Kevin Reynolds	Samuel L. Jackson, Jo...	United States
9	s9	Movie	706	Shravan Kumar	Divya Dutta, Atul Kulka...	India
10	s10	Movie	1920	Vikram Bhatt	Rajneesh Duggal, Ada...	India

Toplam Kayıt Sayısı: 7.787

Veri setimizin değişkenleri(Sütunları);

Tüm veri kümelerinde arama yap	
Alan ara	
netflix_titles	
show_id	
type	
title	
director	
cast	
country	
date_added	
release_year	
rating	
duration	
listed_in	
description	

Ver keşfine değişken bazında devam edelim;

-Tablo üzerindeki değişkenlerin yanındaki ok işaretine baktığımızda;

↑ type	↕ title	↕ direct
TV Show		
Movie		
Movie		
Movie		
Movie		
TV Show		
Movie		
Movie	187	Kevin Re
Movie	706	Shravan
Movie	1920	Vikram B
Movie	1922	Zak Hildit
TV Show	1983	
TV Show	1994	Diego En
Movie	2,215	Nottapon
Movie	3022	John Suit
Movie	Oct-01	Kunle Afc
TV Show	Feb-09	

Aşağıdaki fonksiyonel özelliklerle karşılaşırız

- Veri Türü Dönüştürme
- Değerleri Filtreleme
- Bul ve Değiştir (String değişkenler için)
- Sütunu kaldır
- Sütun özetini göster

Yukarıda ana başlıklarda toplanan fonksiyonel adımların hepsi veri keşfi ,veri özeti ve veri manipülasyonunun çekirdeğini oluşturur.

İsterseniz bu alanların kapsadığı araçları kategorize edelim.

Veri Özeti (Summary) ve Veri Keşfi

- Sütun özetini göster
- Değerleri Filtreleme
- Sütunu Kaldır

Veri Manipülasyonu

- Veri türü dönüştürme
- Bul ve Değiştir

Bu önemli detaydan sonra ilgili araçların içeriklerini inceleyelim.

Veri Türü Dönüştürme:
İlgili değişkenin veri tipini değiştirir.

Veri türünü dönü...

Çıktı veri türü

- Tamsayı
- Çift
- Tarih/Saat

İptal Uygula

type	year	actor
TV Show	1920	Vikram Bhatt
Movie	1922	Zak Hilditch
Movie	1983	
Movie	1994	Diego Enrique Osorno
TV Show	2,215	Nottapon Boonprakob
Movie	3022	John Suits
Movie	Oct-01	Kunle Afolayan
TV Show	Feb-09	

Değerleri Filtrele:

Gözlem değerlerini filtreler.
Type değişkeninde Movie değeri olan kayıtlar filtrelenir.
R'daki karşılığı dplyr paketindeki filter fonksiyonudur.

Değerleri filtrele

Değer ara

☐ Tümünü Seç sayım

☒ Movie 5.377

☐ TV Show 2.410

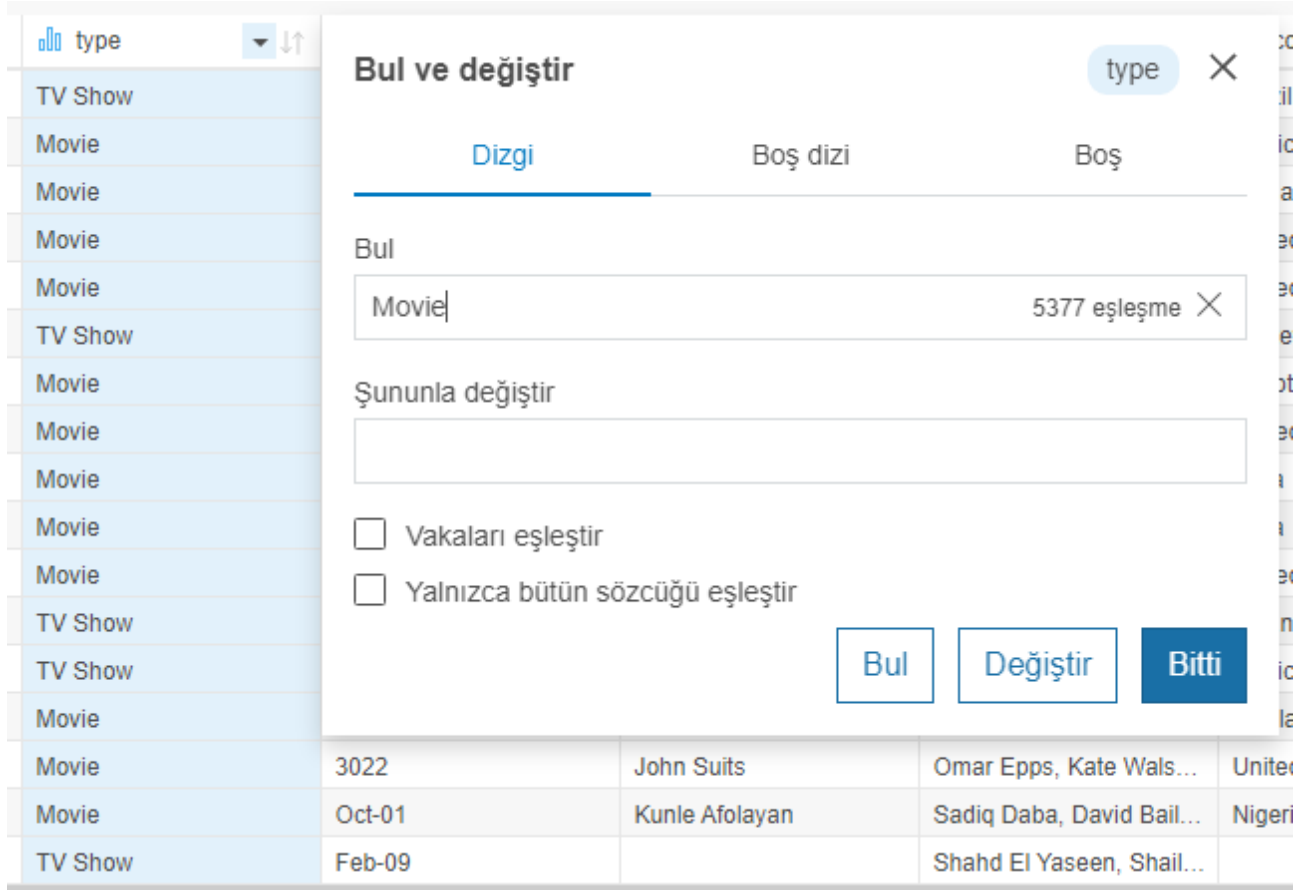
Uygula

type	year	actor
Movie	15-Aug	Swapnaneel Jayakar

Bul ve Değiştir:

Sadece String değişkenlerde çalışır.

R'daki karşılığı replace fonksiyonudur.



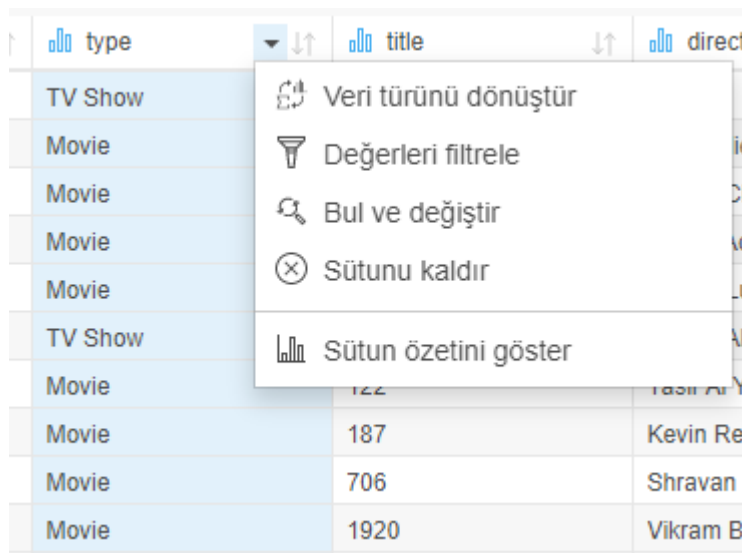
type				
TV Show				
Movie				
Movie				
Movie				
Movie				
TV Show				
Movie				
Movie				
Movie				
Movie				
TV Show				
TV Show				
Movie				
Movie	3022	John Suits	Omar Epps, Kate Wals...	United
Movie	Oct-01	Kunle Afolayan	Sadiq Daba, David Bail...	Nigeri
TV Show	Feb-09		Shahd El Yaseen, Shail...	

Sütunu Kaldır:

tablodan ilgili değişkeni kaldırır.

R'daki karşılığı dplyr paketi select(-c(sütun_ismi)) fonksiyon kombinasyonudur.

Pythondaki karşılığı pandas modülü drop metodudur.



type	title	direct
TV Show		
Movie		
Movie		
Movie		
Movie		
TV Show		
Movie	122	Tasir Ar Y
Movie	187	Kevin Re
Movie	706	Shravan
Movie	1920	Vikram B

Sütun Özetini Göster:

Değişkenin istatistiki özetini gösterir.

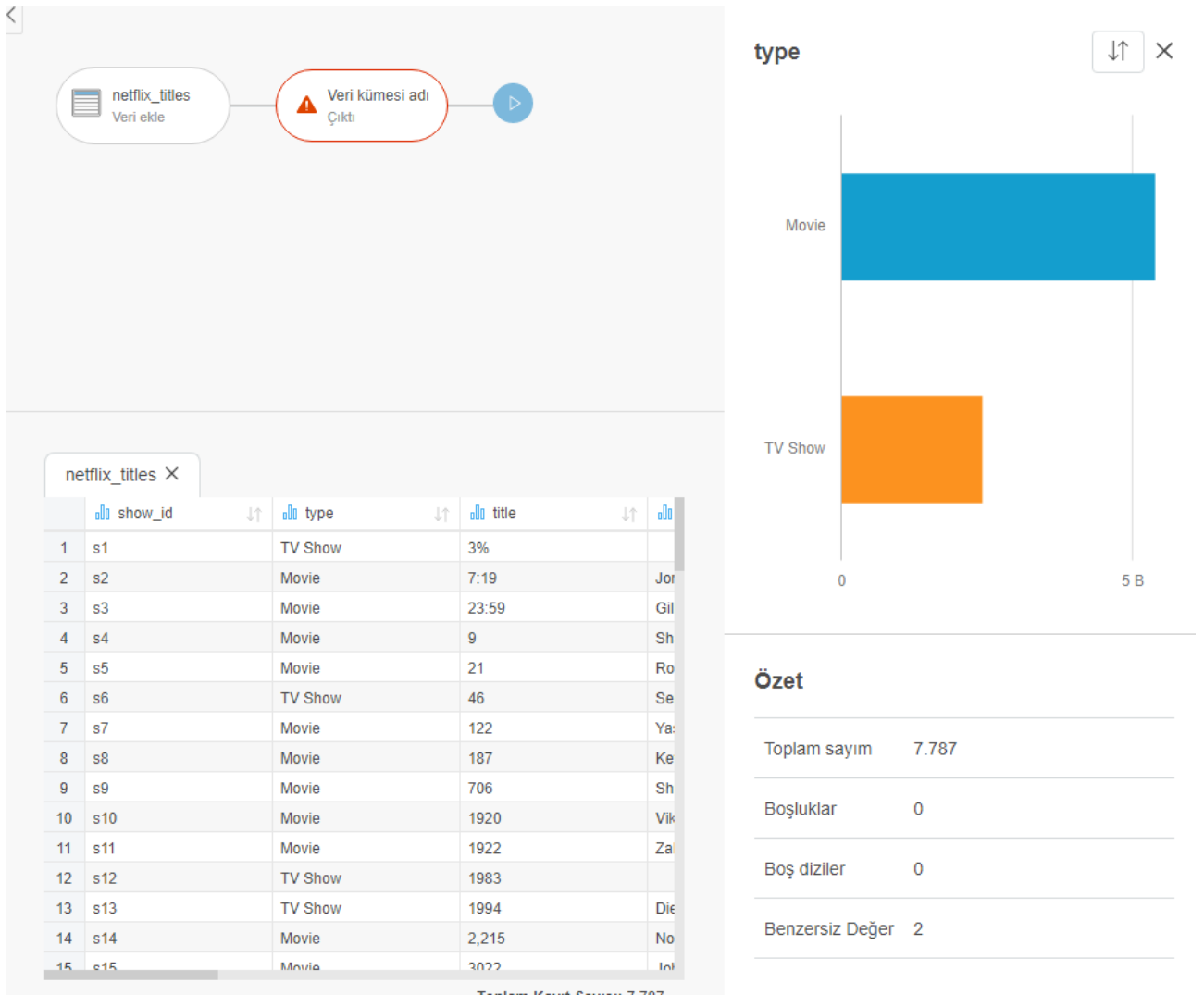
R'daki karşılığı base paketi summary fonksiyonudur veya daha detaylı hali util paketi glmpse fonksiyonudur.

Pythondaki karşılığı describe metodudur.

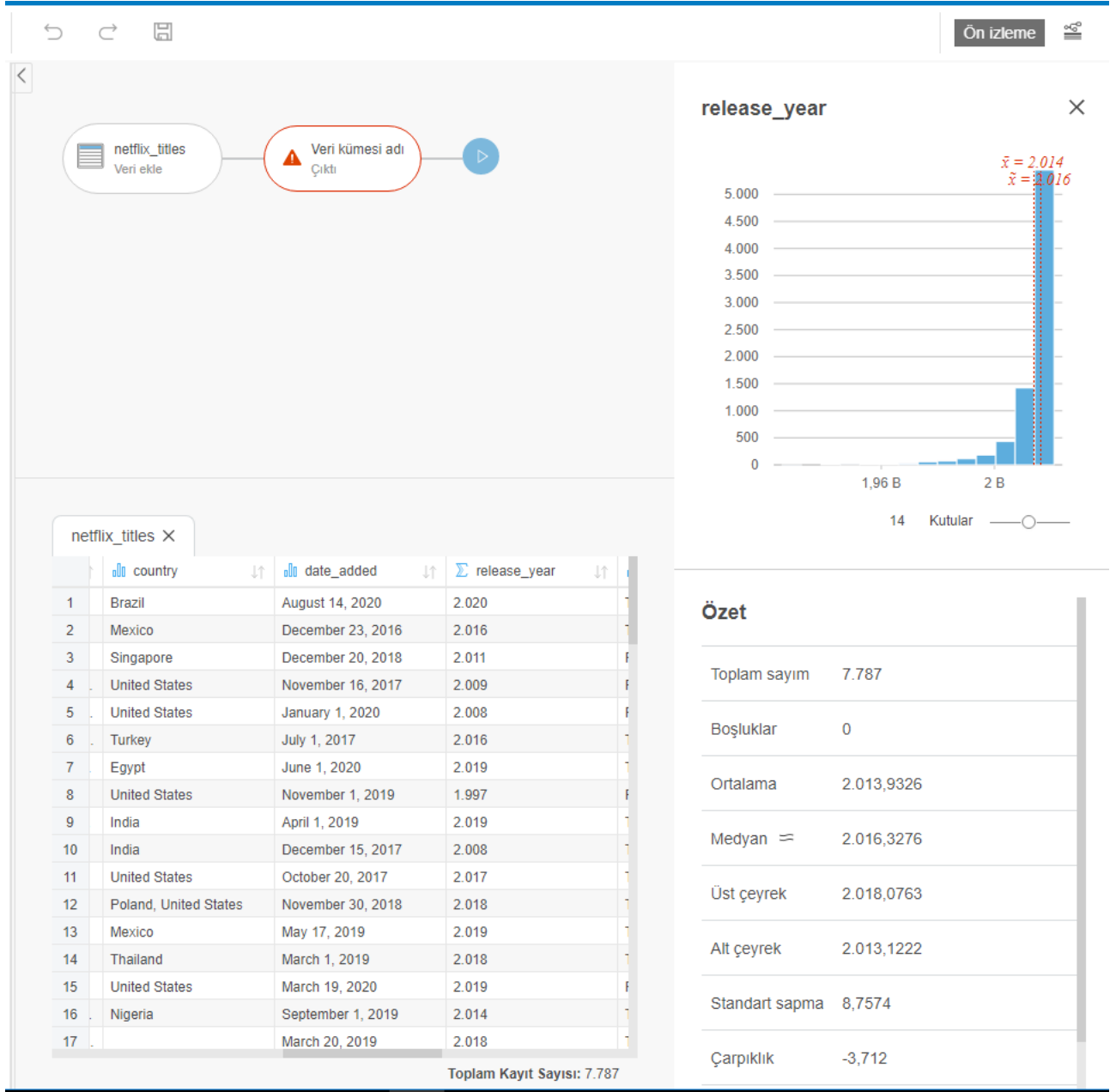
FME Desktop tarafında ise Statistic Calculator ile sağlanır.

Özet tabloda ;

Null değerler,Unique değerler,çeyrek değerleri,Toplam kayıt sayısı gibi değerler özetlenir.



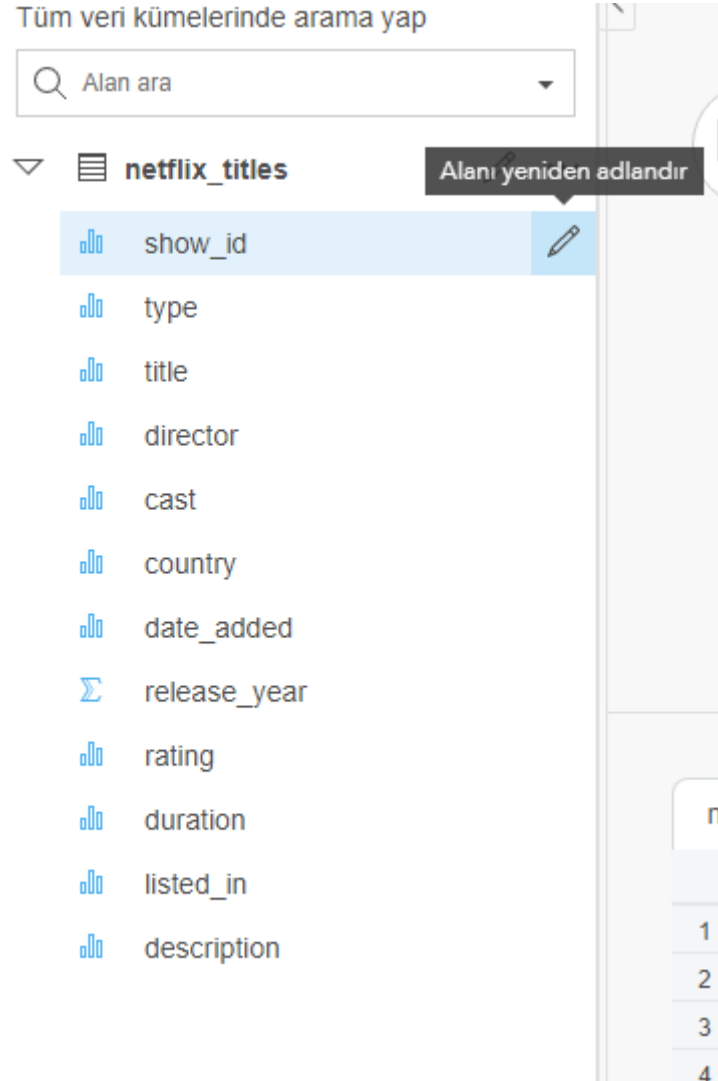
Aşağıda da sayısal bir değişkenin veri özeti görseli bulunuyor.



Öncelikle değişken isimlerine de bakıp değiştirebiliriz.

Bu da R ve Python'da rename fonksiyonuna karşılık gelir.

FME Desktop'ta ise attribute manager transformatörü ile sağlanır.



Veri dağılımını gördükten sonra, yanlış değerleri düzeltmek isteyebiliriz ve bu, **Bul ve değiştir** (Find and replace) aracıyla kolayca yapılabilir. Bu yanlış yazımları, boş değerleri ve boş stringleri değiştirebiliriz.

Aşağıdaki örnekte “director” değişkenindeki “nan” değerleri tespit edip “Boş” değerine dönüştürdük.

director

Bul ve değiştir

director

DizgiBoş diziBoş

Bul

nan76 eşleşme

Şununla değiştir

Boş

☐ Vakaları eşleştir

☐ Yalnızca bütün sözcüğü eşleştir

BulDeğiştirBitti

John Suits	Omar Epps, Kate Wals...	United States	March 19, 2020	2.019
Kunle Afolayan	Sadiq Daba, David Bail...	Nigeria	September 1, 2019	2.014
	Shahd El Yaseen, Shail...		March 20, 2019	2.018

Sütun Manipülasyonu

show_id sütununun değerlerini uppercase edip yeni bir sütuna yazdırıp mevcut show_id sütununu kaldırma işleminin sıralaması aşağıdaki gibidir.

Sütunu hesapl

netflix_titles

Yeni sütun adı

ID

Sütun ifadesi

UPPER(show_id)

Fonksiyonlar

Sütunlar

f_X ABS

f_X AND

f_X AVG

cast

country

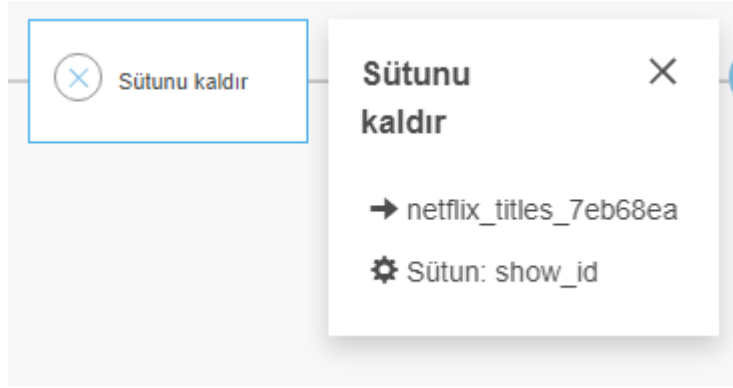
date_added

+ - x ÷ x^y < > = <= >= <> , () AND OR

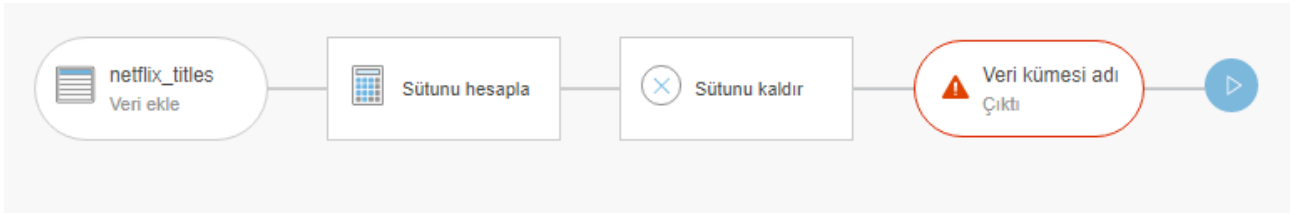
İptal

Çalıştır

show_id sütununu kaldırdık;



modelin genel görünümü ise;



Çıktı verisi için ise ;

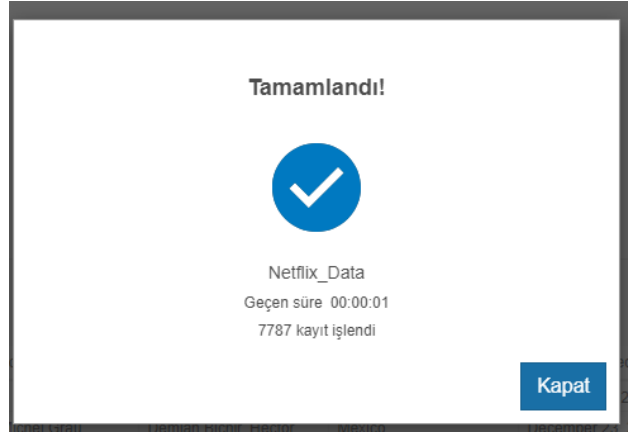
kırmızı işaretli kümeye tıklanır ve karşımıza çıkan pencerede gerekli parametreler doldurulur. Model çalıştırılınca veri yaratılmış olur.

A screenshot of a 'Çıktı konumu' (Output Location) dialog box. The dialog has a sidebar with 'Yerel içerik' and 'Veri Tabanı'. The main area has fields for 'Başlık' (Title), 'Etiket' (Label), and 'Açıklama' (Description). The 'Başlık' field is filled with 'Netflix_Data'. The 'Etiket' field has a button 'Etiket ekle'. The 'Açıklama' field has a placeholder text 'Bu Öğe için bir açıklama girin'. At the bottom, there are 'İptal' and 'Kaydet' buttons.

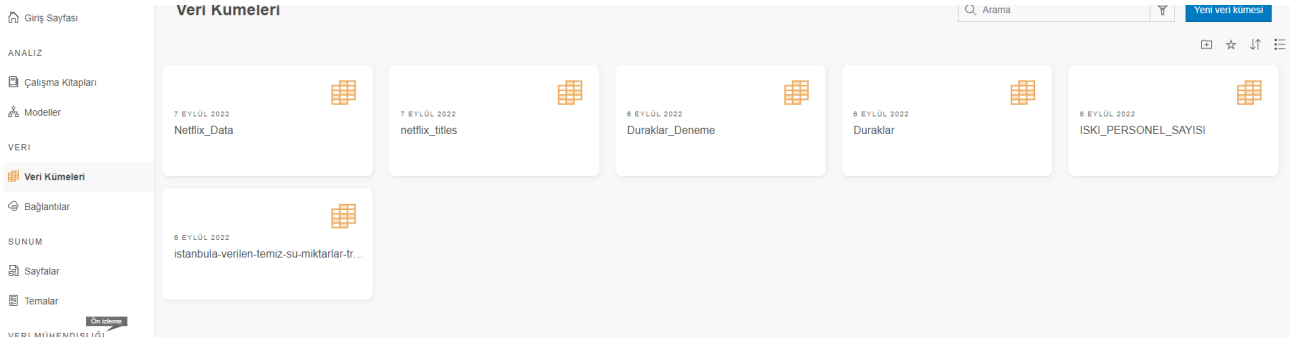
Modelin son hali



Modelin çalışması bittiğinde karşımıza çıkan pencere.



Sonuç ürüne veri kaynaklarımızdan ulaşabiliriz.



Böylelikle veri analize hazır hale gelmiş olur.