BIG Data and Business Intelligence
In-Course Assessment

ACROMEGALY AND IGF ANALYSIS

Section 1 : Business Intelligence Design

15/01/2021

STUDENT

Mohana Kamanooru

A0223038@tees.ac.uk

MODULE LEADER

Dr. Annalisa Occhipinti

a.occhipinti@tees.ac.uk

Acknowledgements

Thanks to my Professor Dr. Annalisa Occhipinti for her continual guidance , support and advice throughout the research process. I want to thank my parents Mr. Ramanachari Kamanooru, Mrs. Vijayalakshmi Kamanooru, and husband, Mr. Santhosh Kanakam, for their continued encouragement and support in my career, and I am very grateful for everything I could learn from all the support received.

Table of Contents

A) Data Source Description and Business Questions

1. Introduction

Human beings are one of the most fantastic creatures in the world. Every single organ and single-cell are essential for human survival. Among the parts of the human body, the pituitary is a tiny gland located behind the bridges of nose attached to the human brain's base. Though the gland's size is small, it is still known as the master gland because it controls all the hormones produced in the human body.

- The problems caused by the pituitary gland are broadly categorized into three types:
- The conditions that alter the size or shape of the gland itself called empty Sella syndrome.
- The conditions which make pituitary to secrete hormone in lower levels than that are required. These are hypopituitarism and diabetes insipidus.

The conditions that cause the pituitary to secrete hormones much more than required like Acromegaly, Cushing's and prolactinoma.

In this thesis, we are interested in Acromegaly, which is a rare pituitary tumour and secretes too much growth hormone GH in the body. The tumour less than 1cm it is called microadenoma, and > 1 cm known as pituitary macroadenoma. They develop DNA mutations and makes cells to grow and divide rapidly. Acromegaly may also result in shortening the life expectancy of the patient. Scientists estimate that about 3 to 14 of every 100,000 people have been diagnosed with Acromegaly. Any research and analysis would be helpful in the medical field, which is snowballing. The DNA, transcript sequence counts, and patient-related data are enormous and complex to analyze or visualize using traditional algorithms and methods. Power BI would work wonders for the same purposes.

2. Purpose of Research

Knowing an extraordinarily little about Acromegaly, my curiosity to understand the disease and its rarity by analyzing in depth encouraged me in choosing this dataset. During the current research I intend to learn the complete process of data analysis, therefore be able to apply these skills systematically to find the required information from the huge data available in real time scenarios.

3. Intended Findings in Study

We analyze the processed data to find if there are any significant **physical differences** between acromegaly patients and Control patients. Also, be able to determine if **age factor** of the patient plays any role in the medical condition. We also study the effects of **IGF** (IGF1, IGF2) and **insulin (blood glucose levels)** levels in both patient categories.

To analyze all the mentioned factors, the data should be properly mapped and the relationships and hidden connections between data should be identified. Then we will be able to choose appropriate visualization tools to present the information drawn from our huge data.

4. Dataset Description

4.1. Data Source

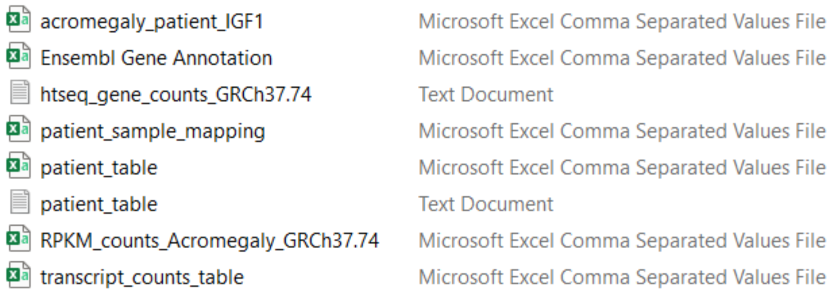
The raw data is captured from the studies carried out by Bridges Lab on neuroendocrine disorders Acromegaly and Cushing's. The raw data is recorded from the patients after clinical and metabolic profiling including HOMA-IR assessment. The physical observations, ceramide levels, insulin glucose, and various other parameters have been recorded in the dataset for patients of both acromegaly and control categories.

4.2. Description

The downloaded dataset contains the raw folder, in which all the patient and sample data is stored in text and CSV files. The file and table information screenshot of the raw folder is shown below.

Table 1 Filenames and Data Tables

No.	File Name	Table Name
Table 1	acromegaly_patient_IGF1.csv	AcromegalyIGF
Table 2	Ensembl Gene Annotation	Ensembl Gene Annotation
Table 3	htseq_gene_counts_GRCh37.74	HTSEQ_Counts
Table 4	patient_sample_mapping	Patient_Sample_Mapping
Table 5	patient_table	Patient_Table
Table 6	RPKM_counts_Acromegaly_GRCh37.74	IGF_RKPM
Table 7	transcript_counts_table	Transcript_Counts



acromegaly_patient_IGF1	Microsoft Excel Comma Separated Values File
Ensembl Gene Annotation	Microsoft Excel Comma Separated Values File
htseq_gene_counts_GRCh37.74	Text Document
patient_sample_mapping	Microsoft Excel Comma Separated Values File
patient_table	Microsoft Excel Comma Separated Values File
patient_table	Text Document
RPKM_counts_Acromegaly_GRCh37.74	Microsoft Excel Comma Separated Values File
transcript_counts_table	Microsoft Excel Comma Separated Values File

Figure 1 Raw Data Files

4.3. Disclosure

This dataset contains the raw data and analysis code for the studies described in this manuscript, publication, and data source links are provided below.

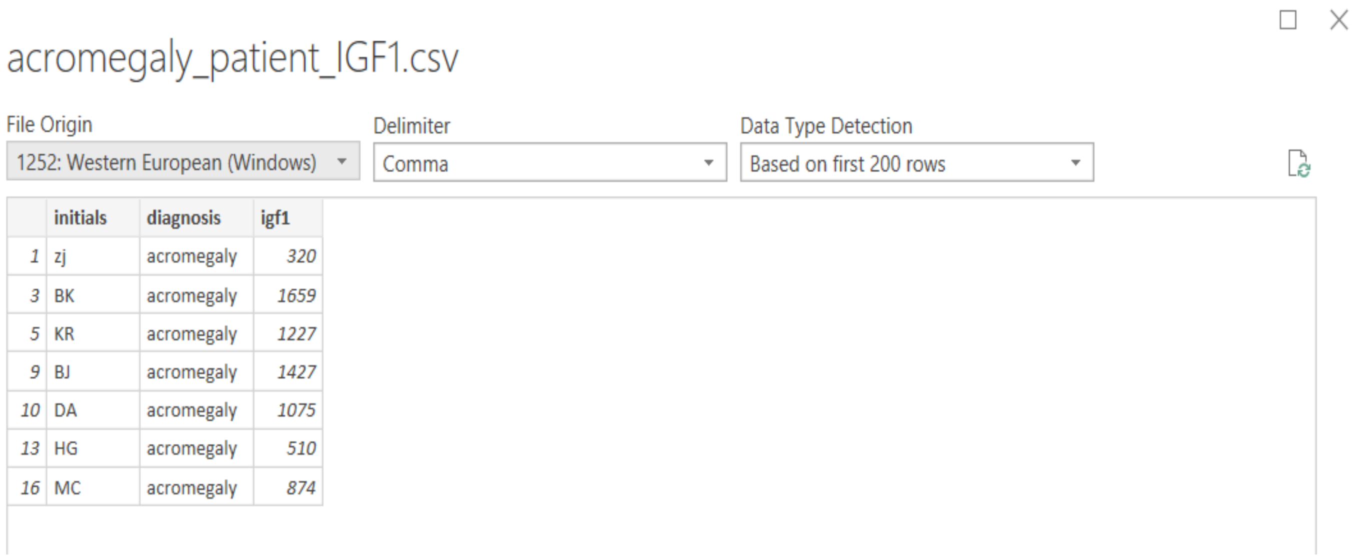
Table 2 Dataset Source

Publication	Dataset	Tag
Hochberg, I, Q. T. Tran, A. L. Barkan, A. R. Saltiel, W. F. Chandler, D. Bridges. Gene Expression Signature in Adipose Tissue of Acromegaly Patients, *PLoS One* 10, e0129359 (2015). doi:10.1371/journal.pone.0129359	Dataset Open Access	Acromegaly-v1.0.0

5. Table Description

5.1 Acromegaly IGF

4 Columns and 8 rows - provides IGF1 levels observed in the patients diagnosed with Acromegaly.



acromegaly_patient_IGF1.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

Data Type Detection: Based on first 200 rows

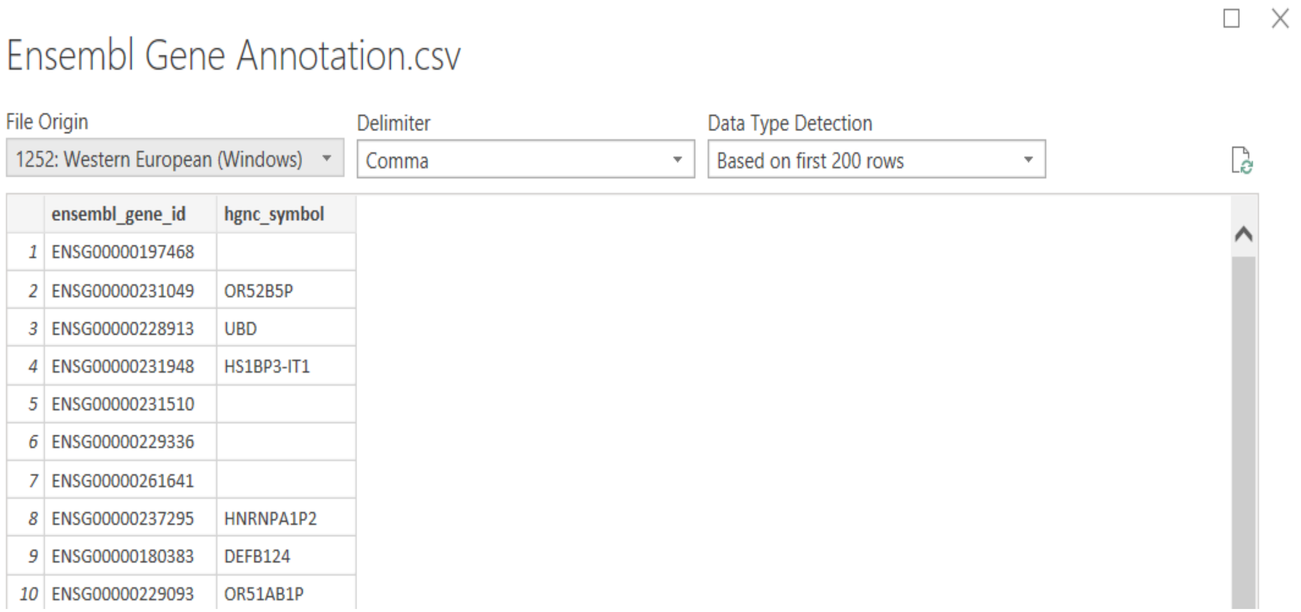
	initials	diagnosis	igf1
1	zj	acromegaly	320
3	BK	acromegaly	1659
5	KR	acromegaly	1227
9	BJ	acromegaly	1427
10	DA	acromegaly	1075
13	HG	acromegaly	510
16	MC	acromegaly	874

Table 3 Acromegaly IGF Columns

Column 1	patient_id	**Id gave to the patients. (identified after analyzing other tables**
Column 2	**patient initials**	First name and Second name Initials of the patient
Column 3	**diagnosis**	Patient’s medical condition
Column 4	**igf1**	levels of IGF1 hormone for respective patients

5.2 Ensembl Gene Annotation

3 Columns and 57383 rows – Provides gene mapping information from Ensembl and HGNC



Ensembl Gene Annotation.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

Data Type Detection: Based on first 200 rows

	ensembl_gene_id	hgnc_symbol
1	ENSG00000197468	
2	ENSG00000231049	OR52B5P
3	ENSG00000228913	UBD
4	ENSG00000231948	HS1BP3-IT1
5	ENSG00000231510	
6	ENSG00000229336	
7	ENSG00000261641	
8	ENSG00000237295	HNRNPA1P2
9	ENSG00000180383	DEFB124
10	ENSG00000229093	OR51AB1P

Table 4 Ensembl Gene Annotation Columns

Column1	index	
Column2	**ensembl_gene_id**	Gene ID from Ensembl Database
Column3	**hgnc_symbol**	Approves gene symbol by HUGO Gene Nomenclature Committee

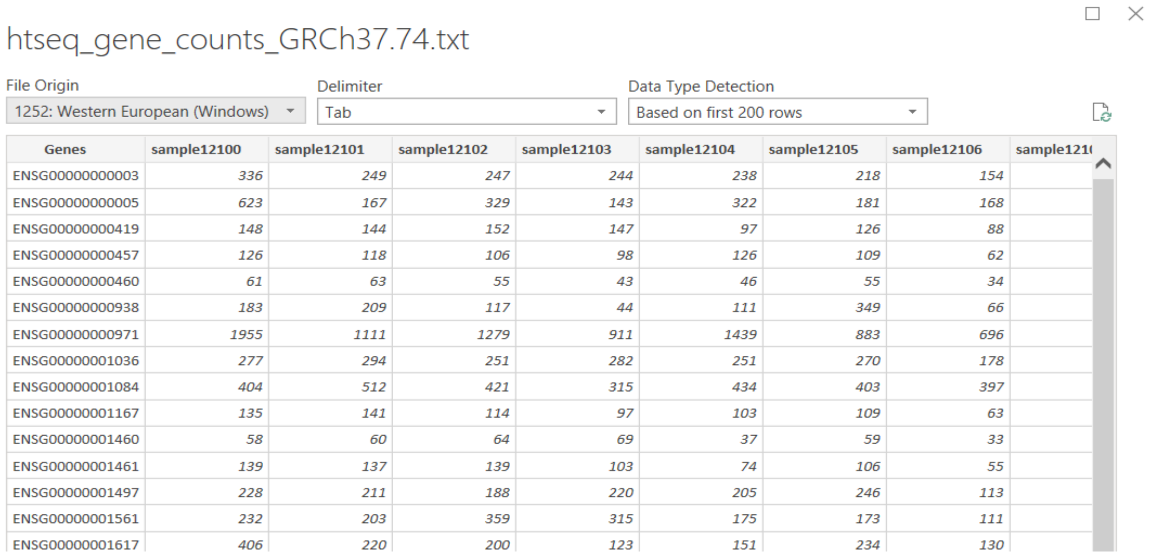
5.3 HTSEQ_Counts

24 Columns and 63684 rows - provides patients gene counts

Table 5 HTSEQ_Counts Columns

Column1	Genes	Gene ID from Ensembl Database

Column2 to 24	Sample121xx	Gene counts for 23 patients respectively



htseq_gene_counts_GRCh37.74.txt

File Origin: 1252: Western European (Windows)

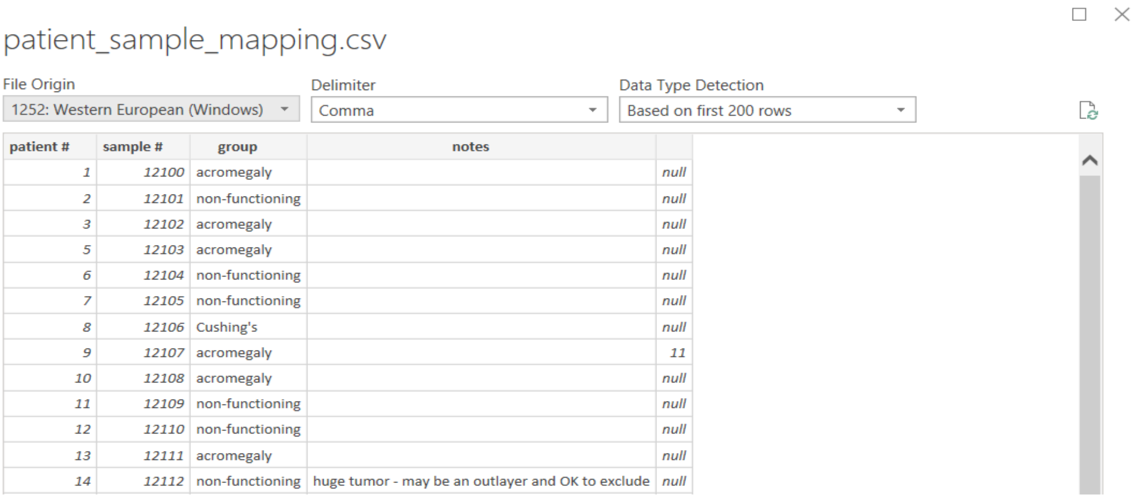
Delimiter: Tab

Data Type Detection: Based on first 200 rows

Genes	sample12100	sample12101	sample12102	sample12103	sample12104	sample12105	sample12106	sample121(
ENSG00000000003	336	249	247	244	238	218	154	
ENSG00000000005	623	167	329	143	322	181	168	
ENSG00000000419	148	144	152	147	97	126	88	
ENSG00000000457	126	118	106	98	126	109	62	
ENSG00000000460	61	63	55	43	46	55	34	
ENSG00000000938	183	209	117	44	111	349	66	
ENSG00000000971	1955	1111	1279	911	1439	883	696	
ENSG00000001036	277	294	251	282	251	270	178	
ENSG00000001084	404	512	421	315	434	403	397	
ENSG00000001167	135	141	114	97	103	109	63	
ENSG00000001460	58	60	64	69	37	59	33	
ENSG00000001461	139	137	139	103	74	106	55	
ENSG00000001497	228	211	188	220	205	246	113	
ENSG00000001561	232	203	359	315	175	173	111	
ENSG00000001617	406	220	200	123	151	234	130	

5.4 Patient_Sample_Mapping

5 columns and 13 rows- This table provides patient and sample mapping information.



patient_sample_mapping.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

Data Type Detection: Based on first 200 rows

patient #	sample #	group	notes	
1	12100	acromegaly		null
2	12101	non-functioning		null
3	12102	acromegaly		null
5	12103	acromegaly		null
6	12104	non-functioning		null
7	12105	non-functioning		null
8	12106	Cushing's		null
9	12107	acromegaly		11
10	12108	acromegaly		null
11	12109	non-functioning		null
12	12110	non-functioning		null
13	12111	acromegaly		null
14	12112	non-functioning	huge tumor - may be an outlayer and OK to exclude	null

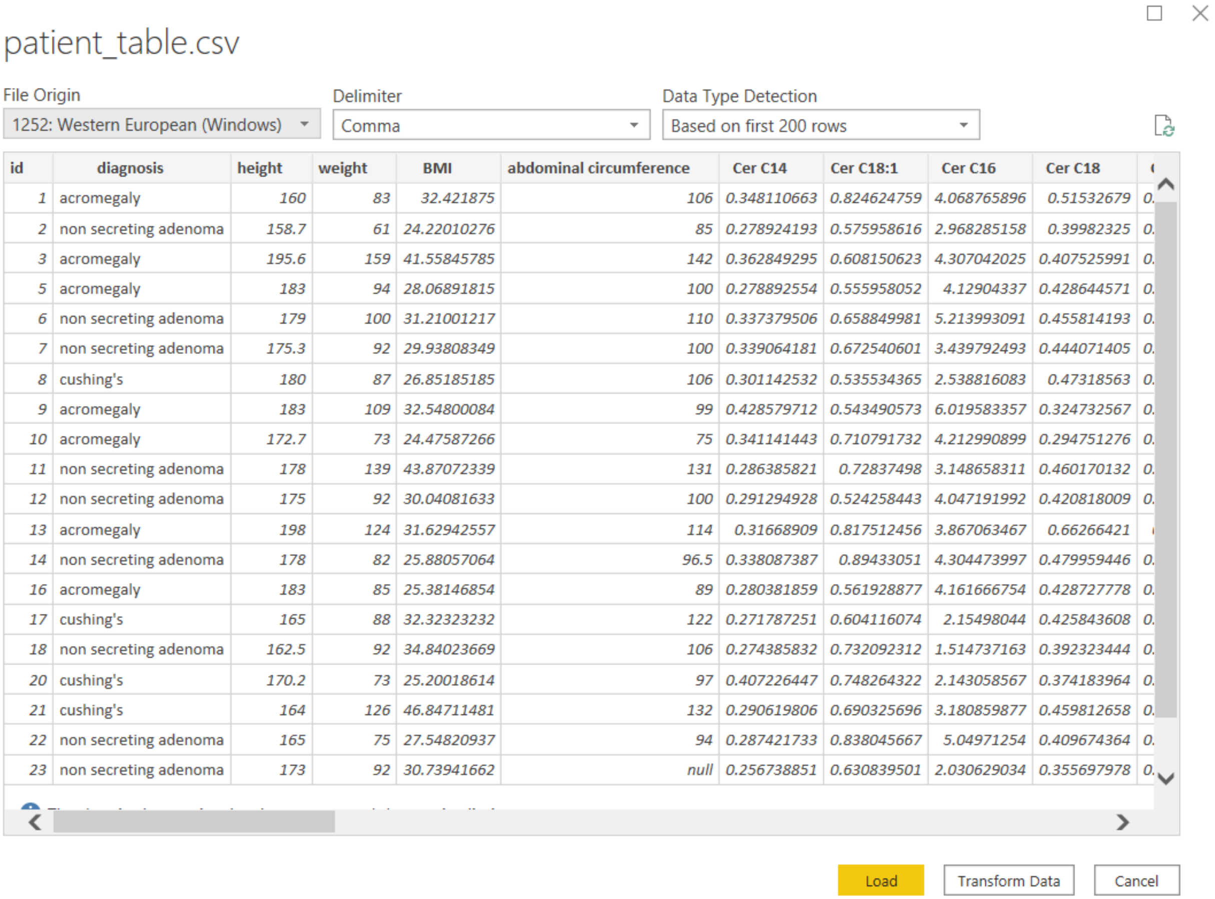
column 4 and column5 have no useful information for analysis.

Table 6 Patient_Sample_Mapping Columns

Column1	Patient_id	patient id
Column2	Sample_id	Gene ID from Ensembl Database
Column3	Group	diagnosis information of the patient.

5.5 Patient Table

36 Columns, 29 rows – Patient observations and details



patient_table.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

Data Type Detection: Based on first 200 rows

id	diagnosis	height	weight	BMI	abdominal circumference	Cer C14	Cer C18:1	Cer C16	Cer C18	
1	acromegaly	160	83	32.421875	106	0.348110663	0.824624759	4.068765896	0.51532679	0.
2	non secreting adenoma	158.7	61	24.22010276	85	0.278924193	0.575958616	2.968285158	0.39982325	0.
3	acromegaly	195.6	159	41.55845785	142	0.362849295	0.608150623	4.307042025	0.407525991	0.
5	acromegaly	183	94	28.06891815	100	0.278892554	0.555958052	4.12904337	0.428644571	0.
6	non secreting adenoma	179	100	31.21001217	110	0.337379506	0.658849981	5.213993091	0.455814193	0.
7	non secreting adenoma	175.3	92	29.93808349	100	0.339064181	0.672540601	3.439792493	0.444071405	0.
8	cushing's	180	87	26.85185185	106	0.301142532	0.535534365	2.538816083	0.47318563	0.
9	acromegaly	183	109	32.54800084	99	0.428579712	0.543490573	6.019583357	0.324732567	0.
10	acromegaly	172.7	73	24.47587266	75	0.341141443	0.710791732	4.212990899	0.294751276	0.
11	non secreting adenoma	178	139	43.87072339	131	0.286385821	0.72837498	3.148658311	0.460170132	0.
12	non secreting adenoma	175	92	30.04081633	100	0.291294928	0.524258443	4.047191992	0.420818009	0.
13	acromegaly	198	124	31.62942557	114	0.31668909	0.817512456	3.867063467	0.66266421	
14	non secreting adenoma	178	82	25.88057064	96.5	0.338087387	0.89433051	4.304473997	0.479959446	0.
16	acromegaly	183	85	25.38146854	89	0.280381859	0.561928877	4.161666754	0.428727778	0.
17	cushing's	165	88	32.32323232	122	0.271787251	0.604116074	2.15498044	0.425843608	0.
18	non secreting adenoma	162.5	92	34.84023669	106	0.274385832	0.732092312	1.514737163	0.392323444	0.
20	cushing's	170.2	73	25.20018614	97	0.407226447	0.748264322	2.143058567	0.374183964	0.
21	cushing's	164	126	46.84711481	132	0.290619806	0.690325696	3.180859877	0.459812658	0.
22	non secreting adenoma	165	75	27.54820937	94	0.287421733	0.838045667	5.04971254	0.409674364	0.
23	non secreting adenoma	173	92	30.73941662	null	0.256738851	0.630839501	2.030629034	0.355697978	0.

Table 7 Patient Table Columns

Column 1	Id	ID allotted to the patient
Column 2	Diagnosis	Patient’s medical condition
Column 3	Height	Height of the patient in cm
Column 4	Weight	Weight of the patient in kg
Column 5	BMI	BMI of the patient in kg/cm2
Column 6	abdominal circumference	Measurement in cm
Column 7	Cer C14	Ceramide species 14:0
Column 8	Cer C18:1	Ceramide species 18:1
Column 9	Cer C16	Ceramide species 16:0
Column 10	Cer C18	Ceramide species 18:0
Column 11	Cer C20	Ceramide species 20:0
Column 12	Cer C22 (area)	Ceramide species 22:0
Column 13	Cer C24:1 (area)	Ceramide species 24:1
Column 14	Cer C24	Ceramide species 24:0
Column 15	Glu-Cer C16	Glucosylcermaide species 16:0
Column 16	Glu-Cer C18	Glucosylcermaide species 18:0
Column 17	Glu-Cer C18:1	Glucosylcermaide species 18:1
Column 18	insulin	Patient’s insulin levels in uIU/ml
Column 19	glucose	Patient’s glucose in mg/dL
Column 20	HOMA-IR	Homeostatic Model Assessment of Insulin Resistance
Column 21	glycerol no tx	Adipose tissue incubation
Column 22	glycerol insulin 2 nM	Adipose tissue incubation with insulin2nM
Column 23	glycerol iso 30 nM	Adipose tissue incubation with isoproterenol 30nM
Column 24	glycerol ins+iso	Adipose tissue incubation with isoproterenol and insulin
Column 25	glycerol ins/ctrl	Adipose tissue incubation with insulin controlled
Column 26	glycerol iso/ctrl	Adipose tissue incubation with isoproterenol controlled
Column 27	glycerol ins+iso/iso	Adipose tissue incubation
Column 28	age	Age of the patient
Column 29	largest diameter of tumor	Size in cm
Column 30	Creatinine	

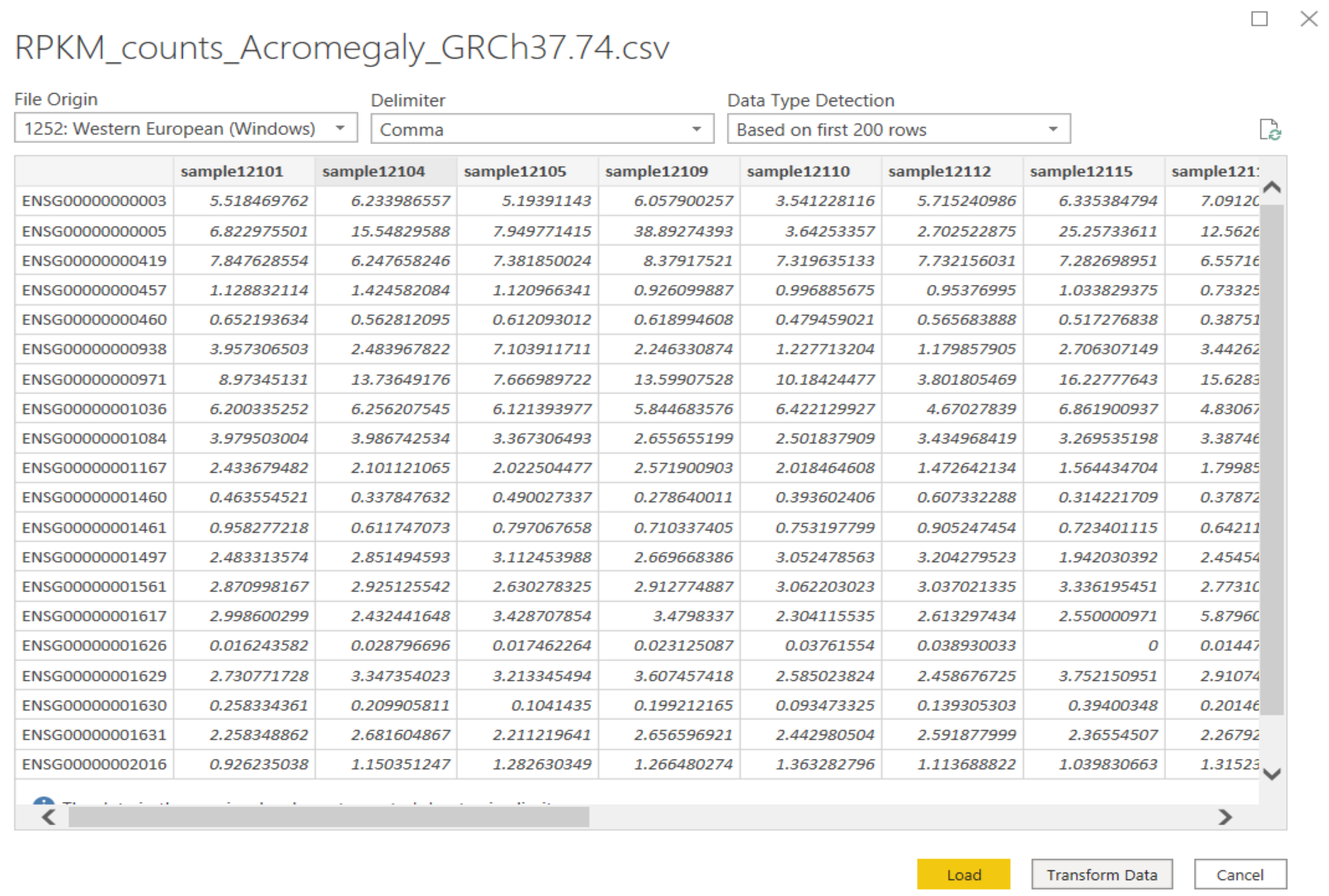
Column 31	AST	
Column 32	ALT	
Column 33	alk phos	
Column 34	HTN	
Column 35	diabetes	If the patient is diabetic or not
Column 36	smoking	Does the patient smoke or not

5.6 IGF_RKPM_Count

RPKM is made for single-end RNA-seq, where every read corresponded to a single fragment that was sequenced.

Table 8 IGF_RKPM_Count Columns

Column1	Genes id	Gene ID from Ensembl Database
Column2 to 24	Sample121xx	Gene counts for 23 patients respectively



RPKM_counts_Acromegaly_GRCh37.74.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

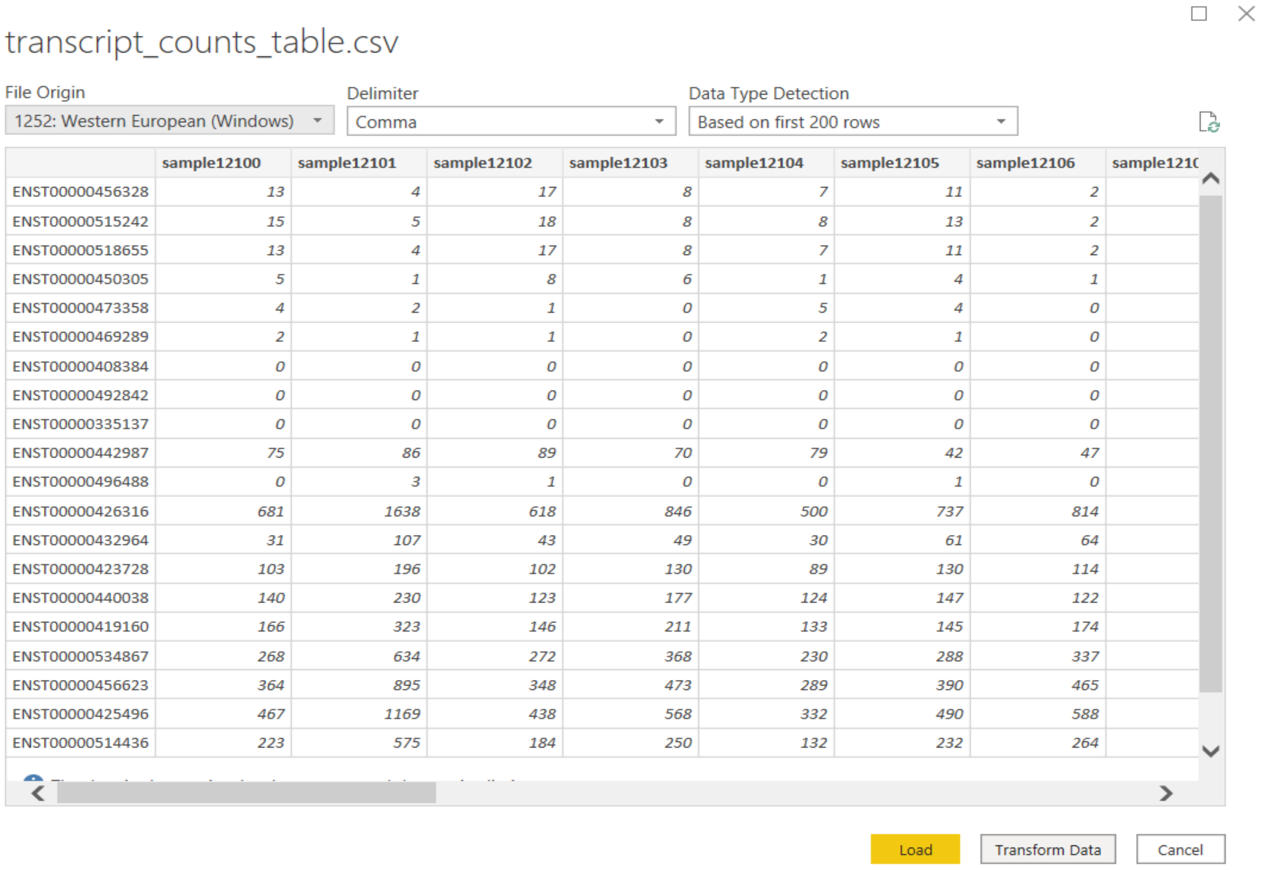
Data Type Detection: Based on first 200 rows

	sample12101	sample12104	sample12105	sample12109	sample12110	sample12112	sample12115	sample121
ENSG00000000003	5.518469762	6.233986557	5.19391143	6.057900257	3.541228116	5.715240986	6.335384794	7.09120
ENSG00000000005	6.822975501	15.54829588	7.949771415	38.89274393	3.64253357	2.702522875	25.25733611	12.5626
ENSG00000000419	7.847628554	6.247658246	7.381850024	8.37917521	7.319635133	7.732156031	7.282698951	6.55716
ENSG00000000457	1.128832114	1.424582084	1.120966341	0.926099887	0.996885675	0.95376995	1.033829375	0.73325
ENSG00000000460	0.652193634	0.562812095	0.612093012	0.618994608	0.479459021	0.565683888	0.517276838	0.38751
ENSG00000000938	3.957306503	2.483967822	7.103911711	2.246330874	1.227713204	1.179857905	2.706307149	3.44262
ENSG00000000971	8.97345131	13.73649176	7.666989722	13.59907528	10.18424477	3.801805469	16.22777643	15.6283
ENSG00000001036	6.200335252	6.256207545	6.121393977	5.844683576	6.422129927	4.67027839	6.861900937	4.83067
ENSG00000001084	3.979503004	3.986742534	3.367306493	2.655655199	2.501837909	3.434968419	3.269535198	3.38746
ENSG00000001167	2.433679482	2.101121065	2.022504477	2.571900903	2.018464608	1.472642134	1.564434704	1.79985
ENSG00000001460	0.463554521	0.337847632	0.490027337	0.278640011	0.393602406	0.607332288	0.314221709	0.37872
ENSG00000001461	0.958277218	0.611747073	0.797067658	0.710337405	0.753197799	0.905247454	0.723401115	0.64211
ENSG00000001497	2.483313574	2.851494593	3.112453988	2.669668386	3.052478563	3.204279523	1.942030392	2.45454
ENSG00000001561	2.870998167	2.925125542	2.630278325	2.912774887	3.062203023	3.037021335	3.336195451	2.77310
ENSG00000001617	2.998600299	2.432441648	3.428707854	3.4798337	2.304115535	2.613297434	2.550000971	5.87960
ENSG00000001626	0.016243582	0.028796696	0.017462264	0.023125087	0.03761554	0.038930033	0	0.01447
ENSG00000001629	2.730771728	3.347354023	3.213345494	3.607457418	2.585023824	2.458676725	3.752150951	2.91074
ENSG00000001630	0.258334361	0.209905811	0.1041435	0.199212165	0.093473325	0.139305303	0.39400348	0.20146
ENSG00000001631	2.258348862	2.681604867	2.211219641	2.656596921	2.442980504	2.591877999	2.36554507	2.26792
ENSG00000002016	0.926235038	1.150351247	1.282630349	1.266480274	1.363282796	1.113688822	1.039830663	1.31523

5.7 Transcript Count

Table 9 Transcript Count Columns

Column1	Genes	Gene ID from Ensembl Database
Column2 to 24	Sample121xx	Gene transcript counts for 23 patients respectively



transcript_counts_table.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

Data Type Detection: Based on first 200 rows

	sample12100	sample12101	sample12102	sample12103	sample12104	sample12105	sample12106	sample121
ENST00000456328	13	4	17	8	7	11	2	
ENST00000515242	15	5	18	8	8	13	2	
ENST00000518655	13	4	17	8	7	11	2	
ENST00000450305	5	1	8	6	1	4	1	
ENST00000473358	4	2	1	0	5	4	0	
ENST00000469289	2	1	1	0	2	1	0	
ENST00000408384	0	0	0	0	0	0	0	
ENST00000492842	0	0	0	0	0	0	0	
ENST00000335137	0	0	0	0	0	0	0	
ENST00000442987	75	86	89	70	79	42	47	
ENST00000496488	0	3	1	0	0	1	0	
ENST00000426316	681	1638	618	846	500	737	814	
ENST00000432964	31	107	43	49	30	61	64	
ENST00000423728	103	196	102	130	89	130	114	
ENST00000440038	140	230	123	177	124	147	122	
ENST00000419160	166	323	146	211	133	145	174	
ENST00000534867	268	634	272	368	230	288	337	
ENST00000456623	364	895	348	473	289	390	465	
ENST00000425496	467	1169	438	568	332	490	588	
ENST00000514436	223	575	184	250	132	232	264	

B) Data Pre-Processing and Data Cleansing

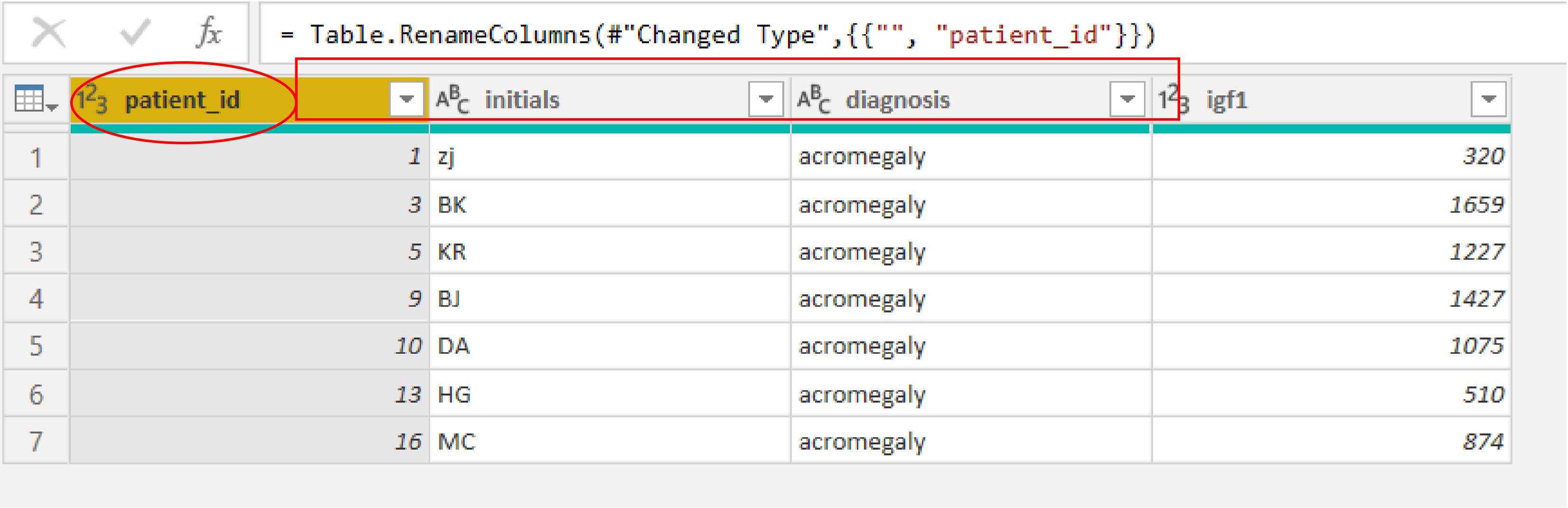
Is there any evidence of steps performed to cleanse the data? For example: • Removing NAs, • Renaming columns • Changing data types • Removing errors • Removing columns • Merging tables, etc

1. Table 1: Acromegaly IGF

1.1. Renaming Columns

Renaming the first blank column to “patient_id” using the M formula shown below.

M Formula = **Table.RenameColumns(#"Changed Type",{{"", "patient_id"}})**



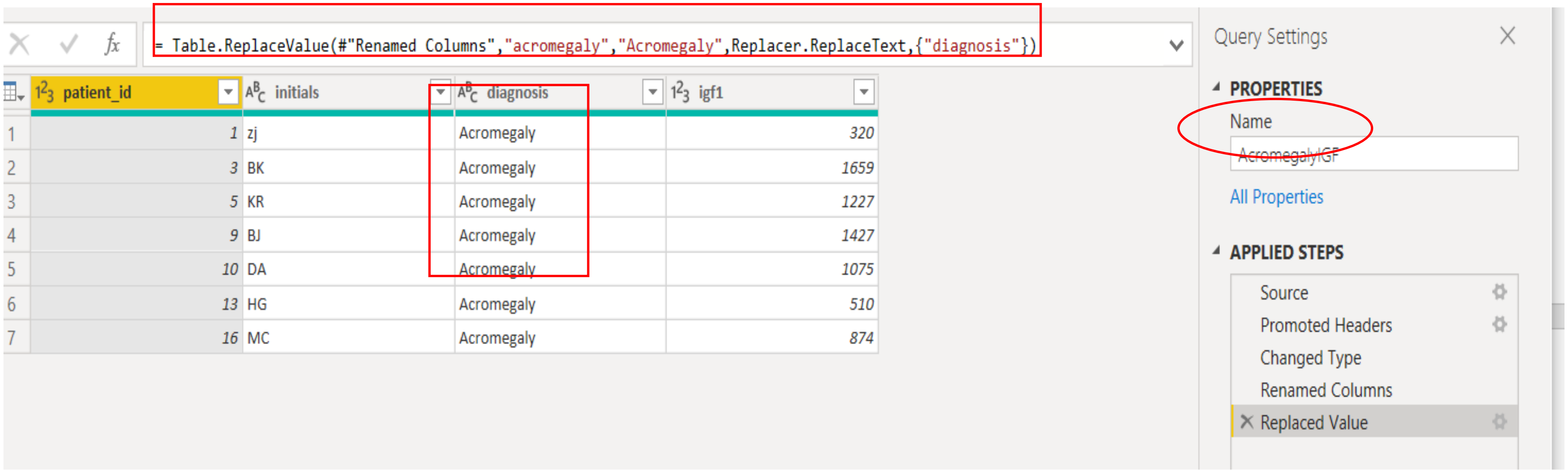
```
= Table.RenameColumns(#"Changed Type",{{"", "patient_id"}})
```

	patient_id	initials	diagnosis	igf1
1	1	zj	acromegaly	320
2	3	BK	acromegaly	1659
3	5	KR	acromegaly	1227
4	9	BJ	acromegaly	1427
5	10	DA	acromegaly	1075
6	13	HG	acromegaly	510
7	16	MC	acromegaly	874

1.2. Replacing Values

The table name is changed to Acromegaly_IGF_Table and replaced diagnosis column values from “acromegaly” to “Acromegaly”.

M formula: = **Table.ReplaceValue(#"Renamed Columns","acromegaly","Acromegaly",Replacer.ReplaceText,{"diagnosis"})**



```
= Table.ReplaceValue(#"Renamed Columns","acromegaly","Acromegaly",Replacer.ReplaceText,{"diagnosis"})
```

	patient_id	initials	diagnosis	igf1
1	1	zj	Acromegaly	320
2	3	BK	Acromegaly	1659
3	5	KR	Acromegaly	1227
4	9	BJ	Acromegaly	1427
5	10	DA	Acromegaly	1075
6	13	HG	Acromegaly	510
7	16	MC	Acromegaly	874

Query Settings

PROPERTIES

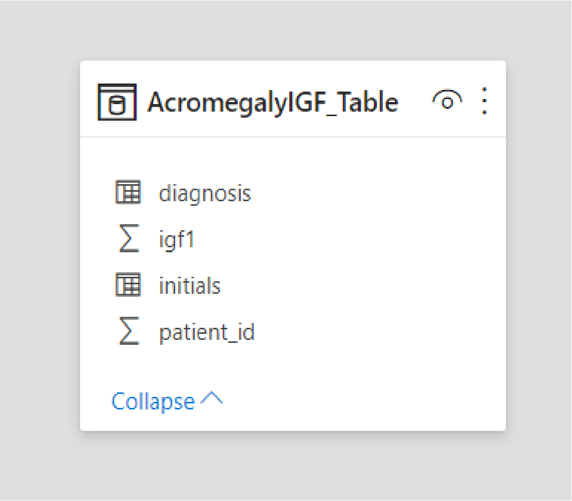
Name: AcromegalyIGF

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Renamed Columns
- Replaced Value

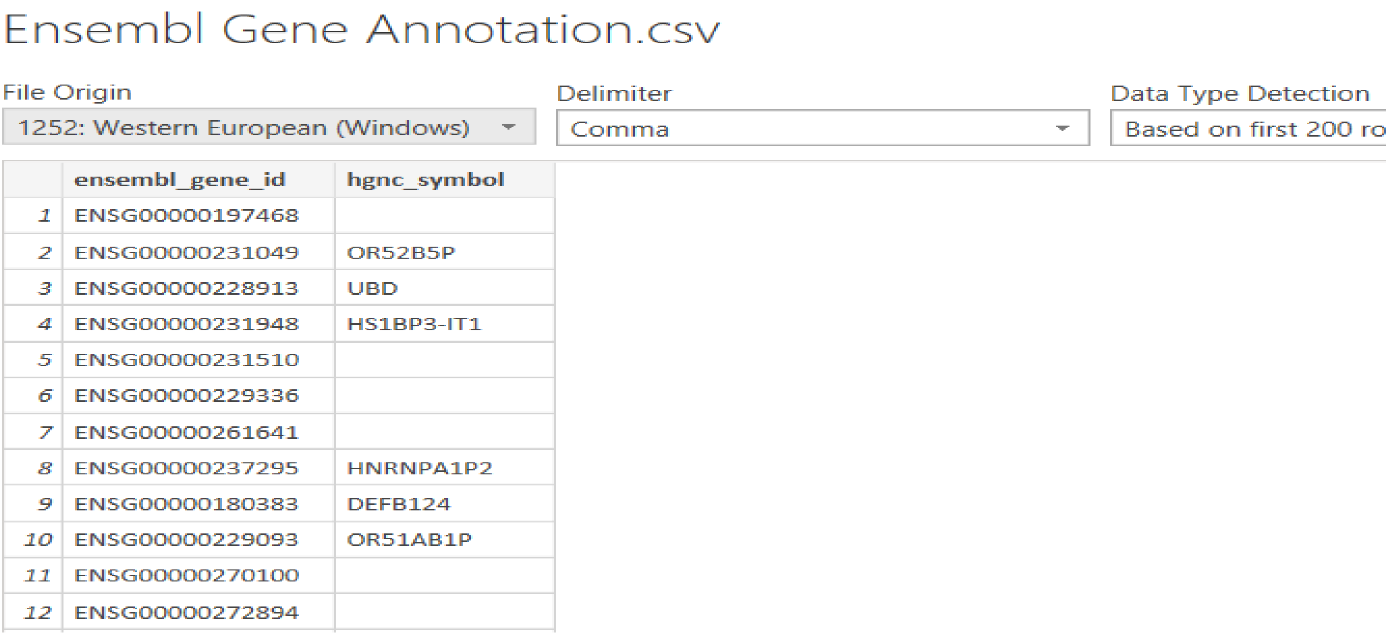
1.3. Model View



AcromegalyIGF_Table

- diagnosis
- ∑ igf1
- initials
- ∑ patient_id

2. Table 2: Ensembl Gene Annotation



Ensembl Gene Annotation.csv

File Origin: 1252: Western European (Windows)

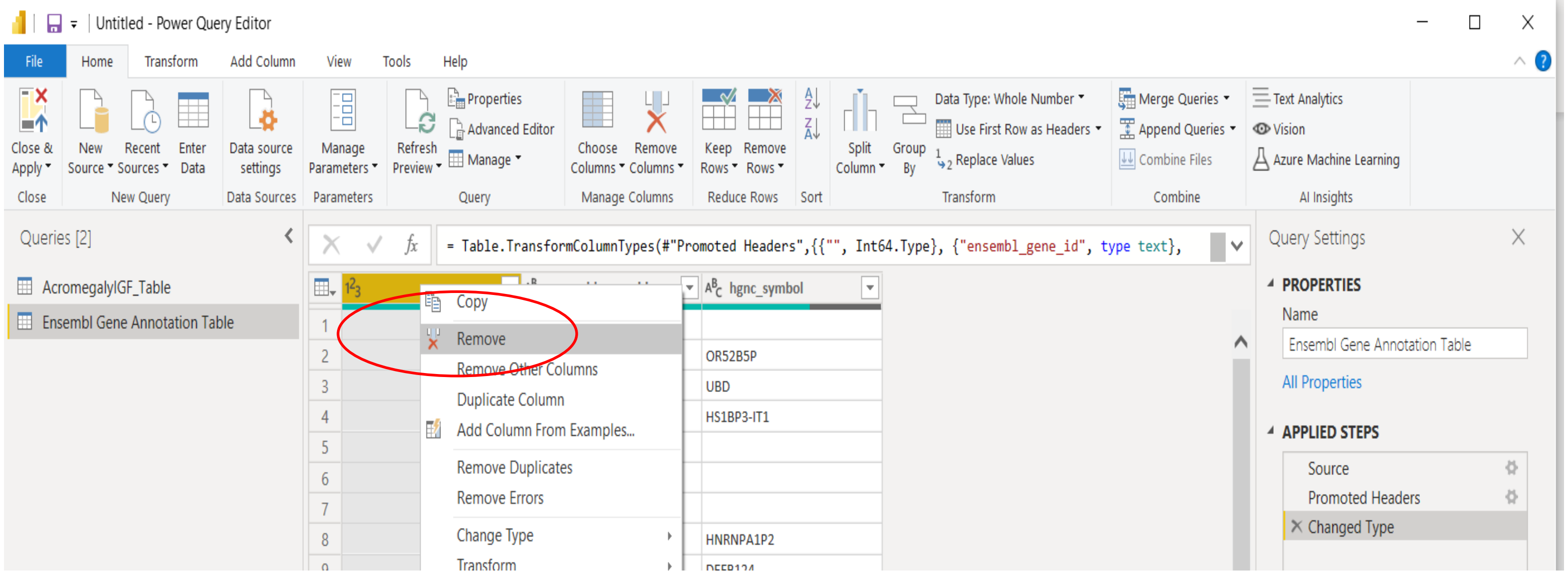
Delimiter: Comma

Data Type Detection: Based on first 200 ro

	ensembl_gene_id	hgnc_symbol
1	ENSG00000197468	
2	ENSG00000231049	OR52B5P
3	ENSG00000228913	UBD
4	ENSG00000231948	HS1BP3-IT1
5	ENSG00000231510	
6	ENSG00000229336	
7	ENSG00000261641	
8	ENSG00000237295	HNRNPA1P2
9	ENSG00000180383	DEFB124
10	ENSG00000229093	OR51AB1P
11	ENSG00000270100	
12	ENSG00000272894	

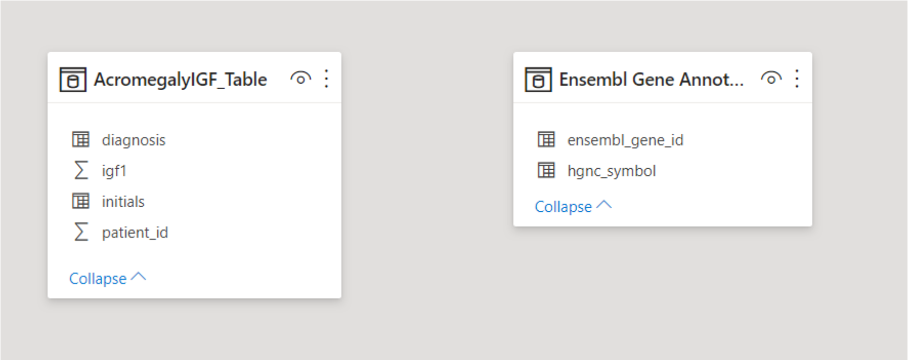
2.1. Removing Columns

The first blank column has index values and is not very useful in analysis. Removing the column as below.

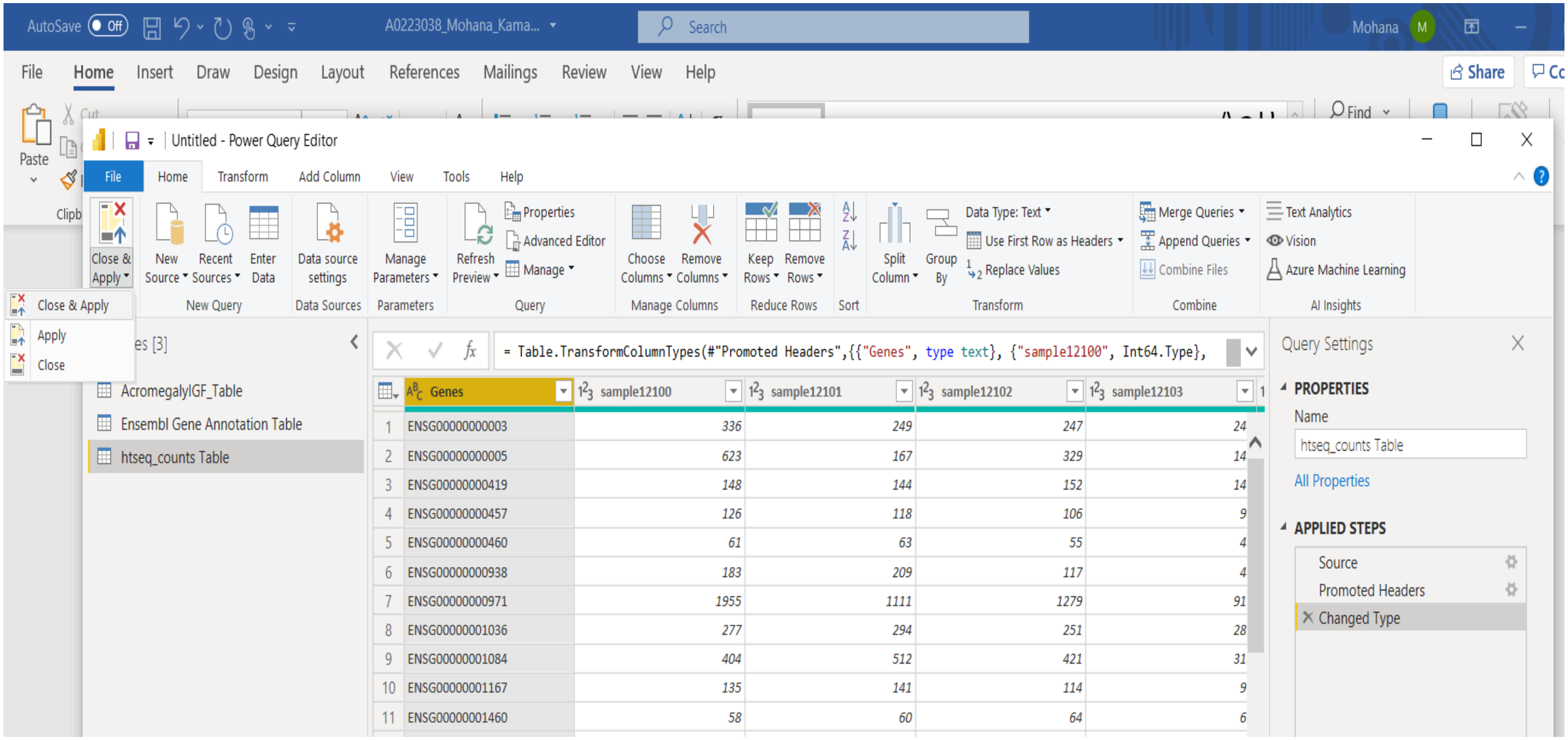


```
= Table.TransformColumnTypes(#"Promoted Headers",{{"", Int64.Type}, {"ensembl_gene_id", type text},
```

2.2. Model View



3. Table 3: HTSEQ_Counts

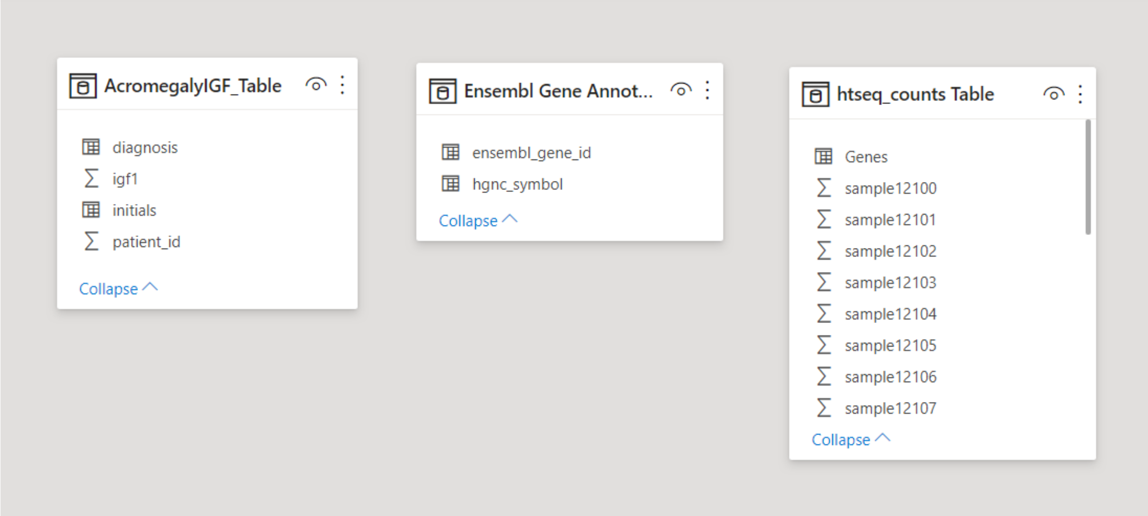


```
= Table.TransformColumnTypes(#"Promoted Headers",{{"Genes", type text}, {"sample12100", Int64.Type},
```

	Genes	sample12100	sample12101	sample12102	sample12103
1	ENSG00000000003	336	249	247	24
2	ENSG00000000005	623	167	329	14
3	ENSG00000000419	148	144	152	14
4	ENSG00000000457	126	118	106	9
5	ENSG00000000460	61	63	55	4
6	ENSG00000000938	183	209	117	4
7	ENSG00000000971	1955	1111	1279	91
8	ENSG00000001036	277	294	251	28
9	ENSG00000001084	404	512	421	31
10	ENSG00000001167	135	141	114	9
11	ENSG00000001460	58	60	64	6

Name: htseq_counts Table

3.1. Model View



AcromegalyIGF_Table

- diagnosis
- igf1
- initials
- patient_id

Collapse

Ensembl Gene Annot...

- ensembl_gene_id
- hgnc_symbol

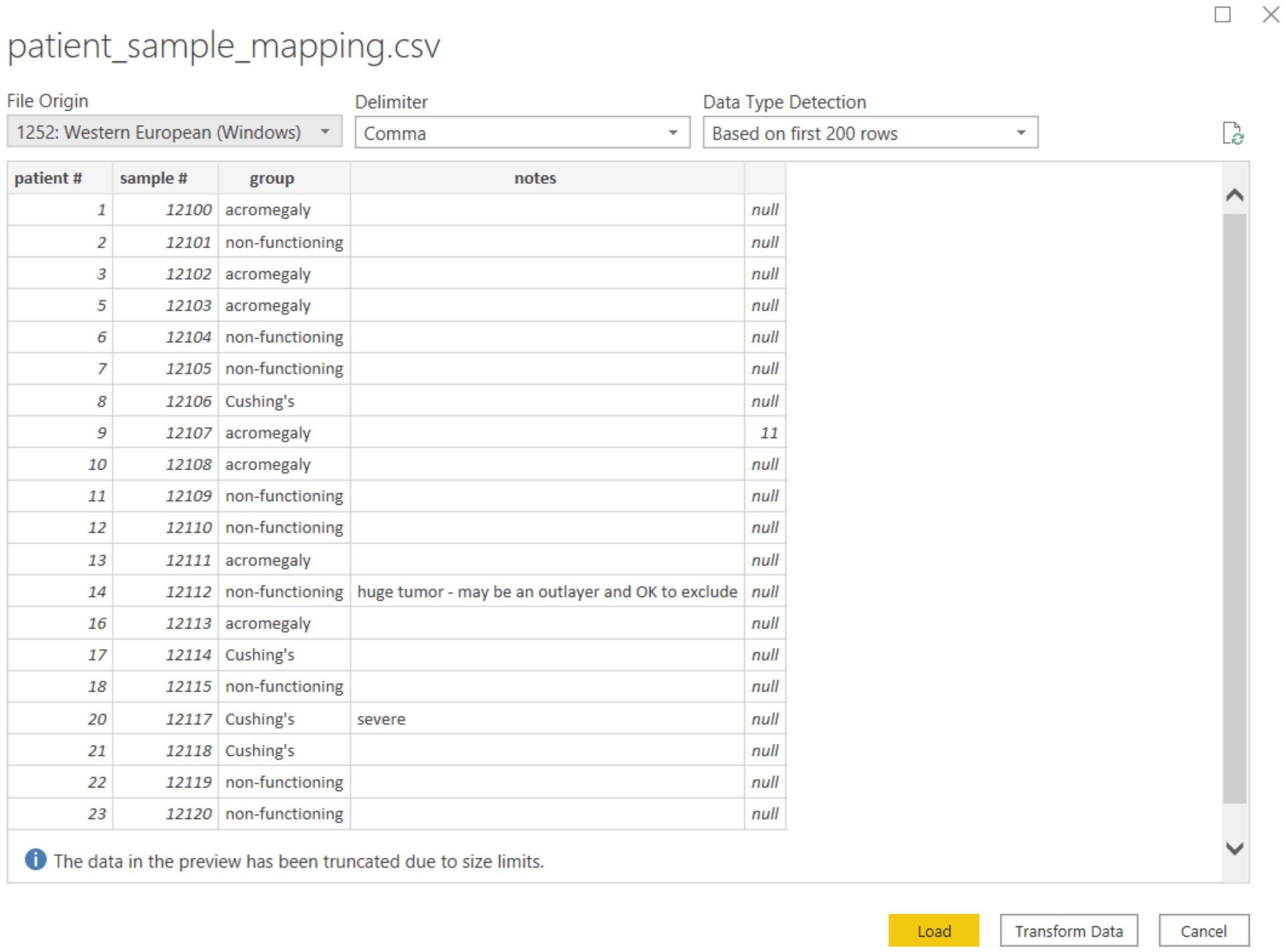
Collapse

htseq_counts Table

- Genes
- sample12100
- sample12101
- sample12102
- sample12103
- sample12104
- sample12105
- sample12106
- sample12107

Collapse

4. Table 4: Patient_sample_maping



patient_sample_mapping.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

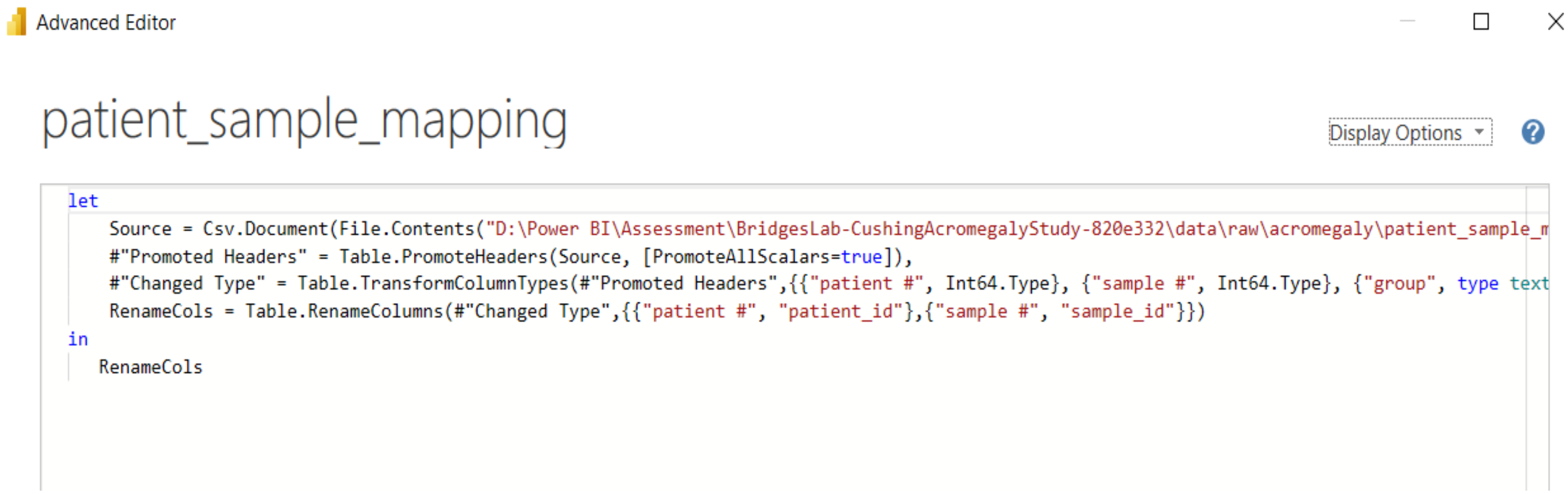
Data Type Detection: Based on first 200 rows

patient #	sample #	group	notes	
1	12100	acromegaly		null
2	12101	non-functioning		null
3	12102	acromegaly		null
5	12103	acromegaly		null
6	12104	non-functioning		null
7	12105	non-functioning		null
8	12106	Cushing's		null
9	12107	acromegaly		11
10	12108	acromegaly		null
11	12109	non-functioning		null
12	12110	non-functioning		null
13	12111	acromegaly		null
14	12112	non-functioning	huge tumor - may be an outlayer and OK to exclude	null
16	12113	acromegaly		null
17	12114	Cushing's		null
18	12115	non-functioning		null
20	12117	Cushing's	severe	null
21	12118	Cushing's		null
22	12119	non-functioning		null
23	12120	non-functioning		null

The data in the preview has been truncated due to size limits.

4.1. Renaming Columns

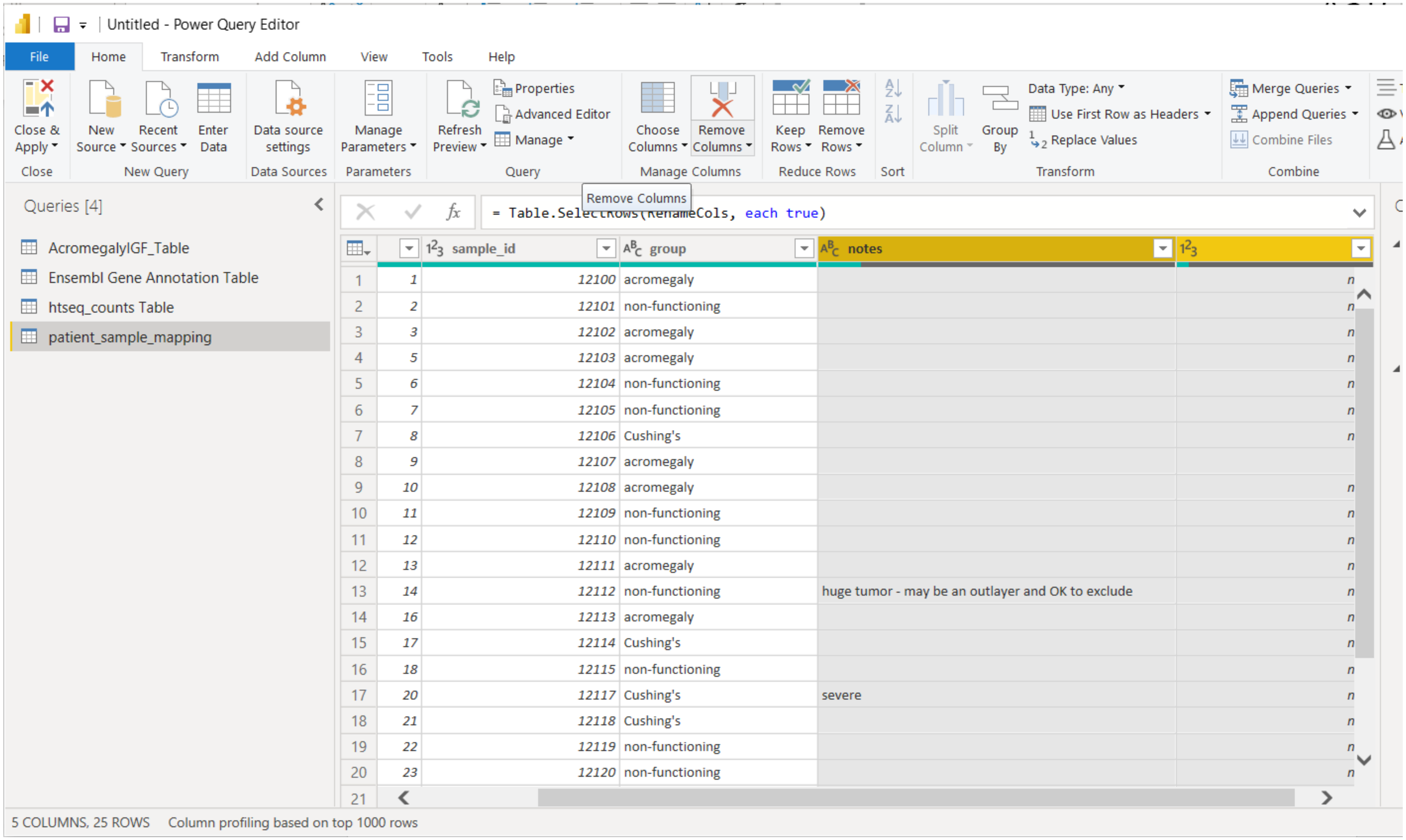
Renaming the first two columns of the table from “patient #” to “patient_id” and “sample #” to “sample_id” using M language in the advanced editor.



```
let
    Source = Csv.Document(File.Contents("D:\Power BI\Assessment\BridgesLab-CushingAcromegalyStudy-820e332\data\raw\acromegaly\patient_sample_m
    #"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"patient #", Int64.Type}, {"sample #", Int64.Type}, {"group", type text
    RenameCols = Table.RenameColumns(#"Changed Type",{{"patient #", "patient_id"},{"sample #", "sample_id"}})
in
   RenameCols
```

4.2. Removing Columns

The last two columns “notes” and the blank column at the end do not have useful information for analysis. Delete the two columns highlighted below.



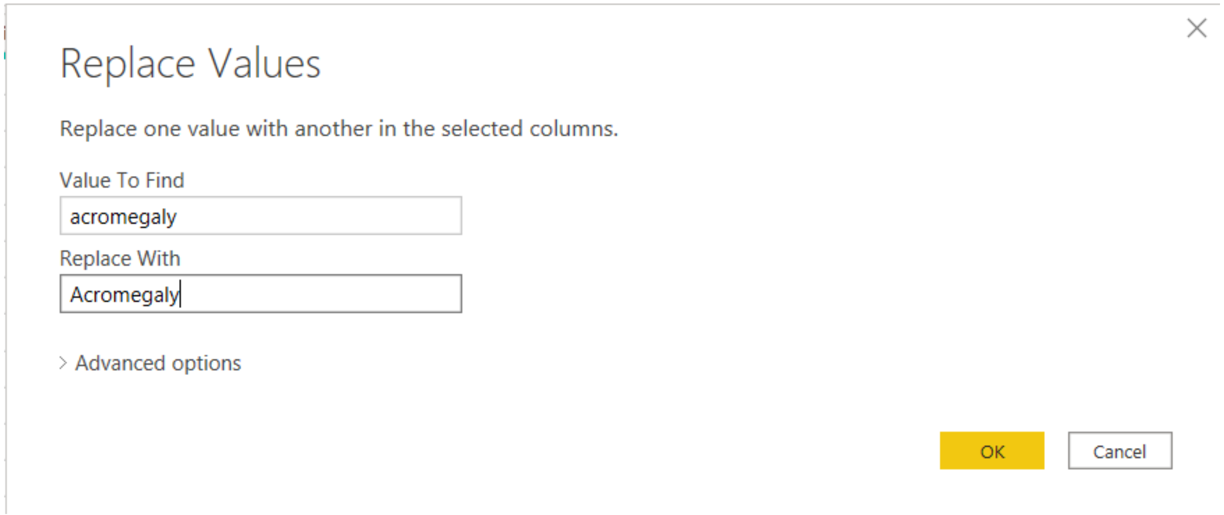
```
= Table.SelectRows(RenameCols, each true)
```

		sample_id	group	notes	
1	1	12100	acromegaly		n
2	2	12101	non-functioning		n
3	3	12102	acromegaly		n
4	5	12103	acromegaly		n
5	6	12104	non-functioning		n
6	7	12105	non-functioning		n
7	8	12106	Cushing's		n
8	9	12107	acromegaly		
9	10	12108	acromegaly		n
10	11	12109	non-functioning		n
11	12	12110	non-functioning		n
12	13	12111	acromegaly		n
13	14	12112	non-functioning	huge tumor - may be an outlayer and OK to exclude	n
14	16	12113	acromegaly		n
15	17	12114	Cushing's		n
16	18	12115	non-functioning		n
17	20	12117	Cushing's	severe	n
18	21	12118	Cushing's		n
19	22	12119	non-functioning		n
20	23	12120	non-functioning		n
21					

5 COLUMNS, 25 ROWS Column profiling based on top 1000 rows

4.3. Replacing Values

Replace the “group “column values from “acromegaly” to “Acromegaly “



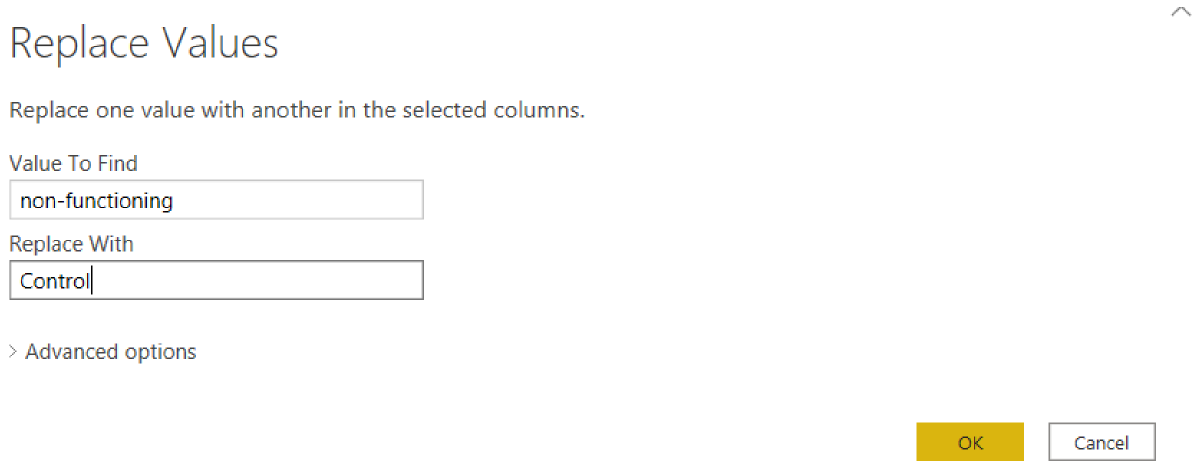
Replace Values

Replace one value with another in the selected columns.

Value To Find: acromegaly

Replace With: Acromegaly

Also, replace “non-functioning” with Control



Replace Values

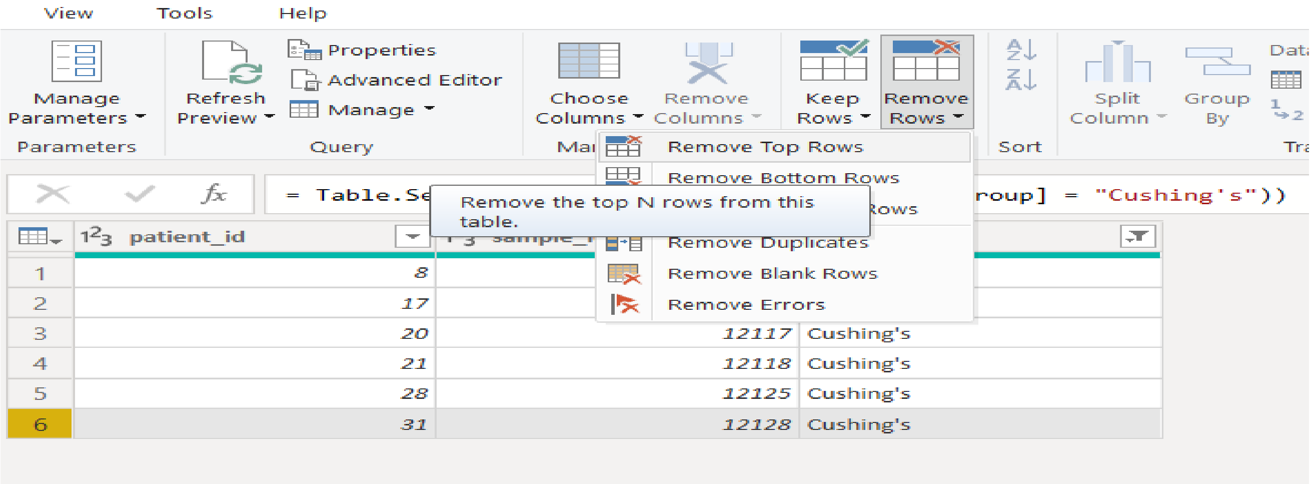
Replace one value with another in the selected columns.

Value To Find: non-functioning

Replace With: Control

4.4. Filtering Rows

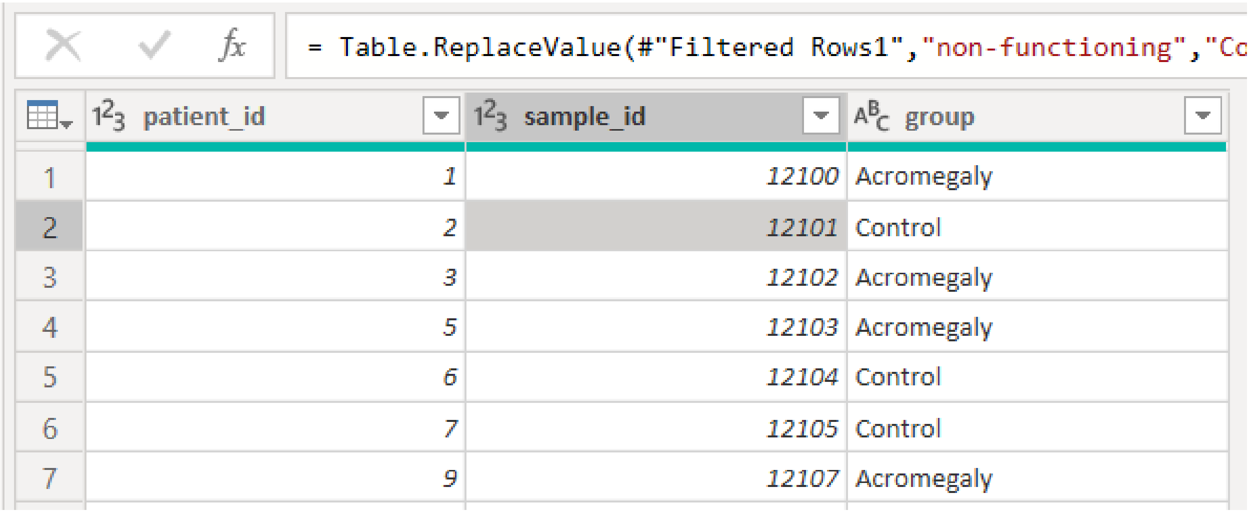
The dataset consists of data related to Acromegaly and Cushing’s and Normal patients. The analysis is based only on Acromegaly and Control Patients. Hence filtering the Cushing’s columns from the table.



Remove the top N rows from this table.

```
= Table.Se                                  roup] = "Cushing's"))
```

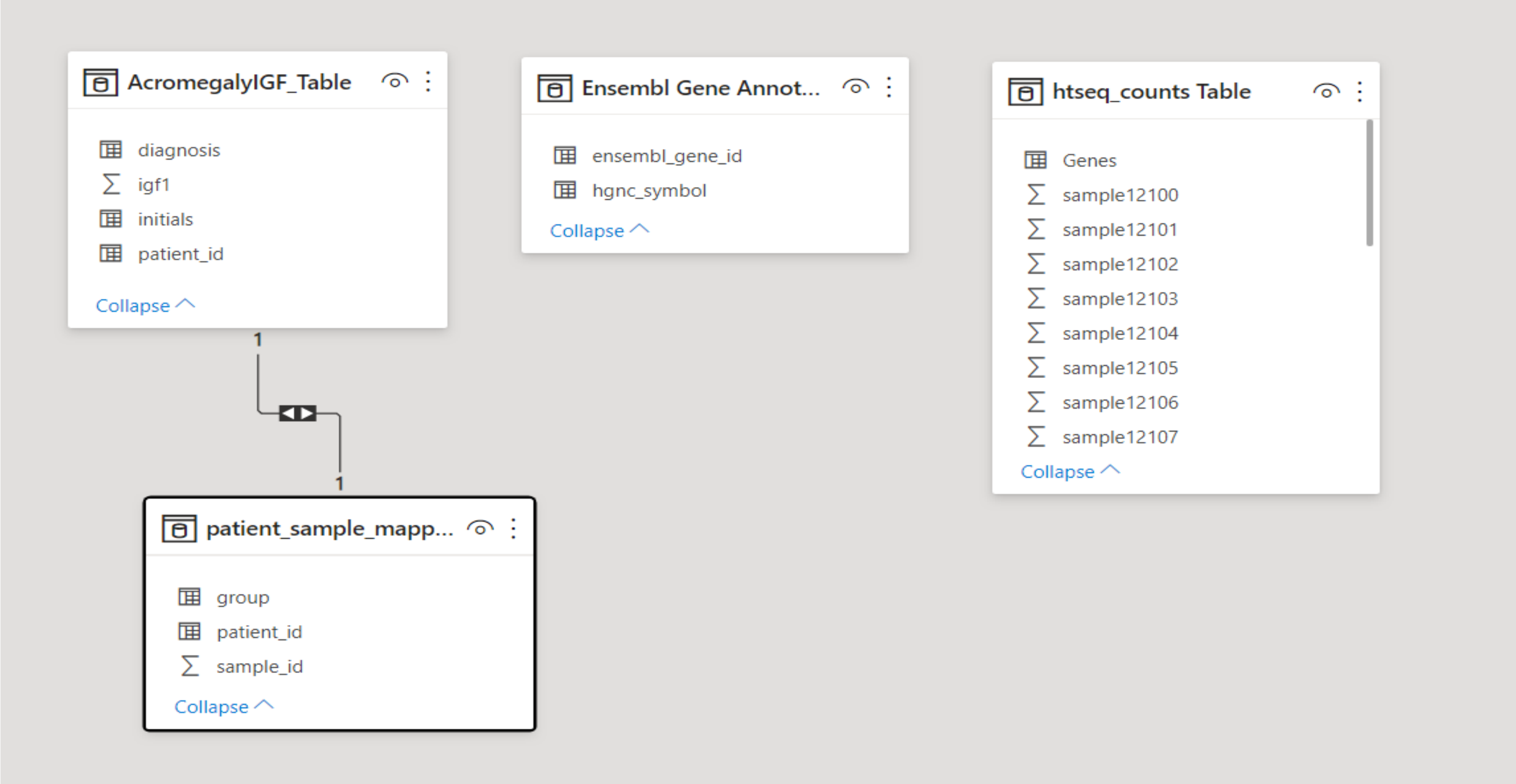
	patient_id	sample_	
1	8		
2	17		
3	20	12117	Cushing's
4	21	12118	Cushing's
5	28	12125	Cushing's
6	31	12128	Cushing's



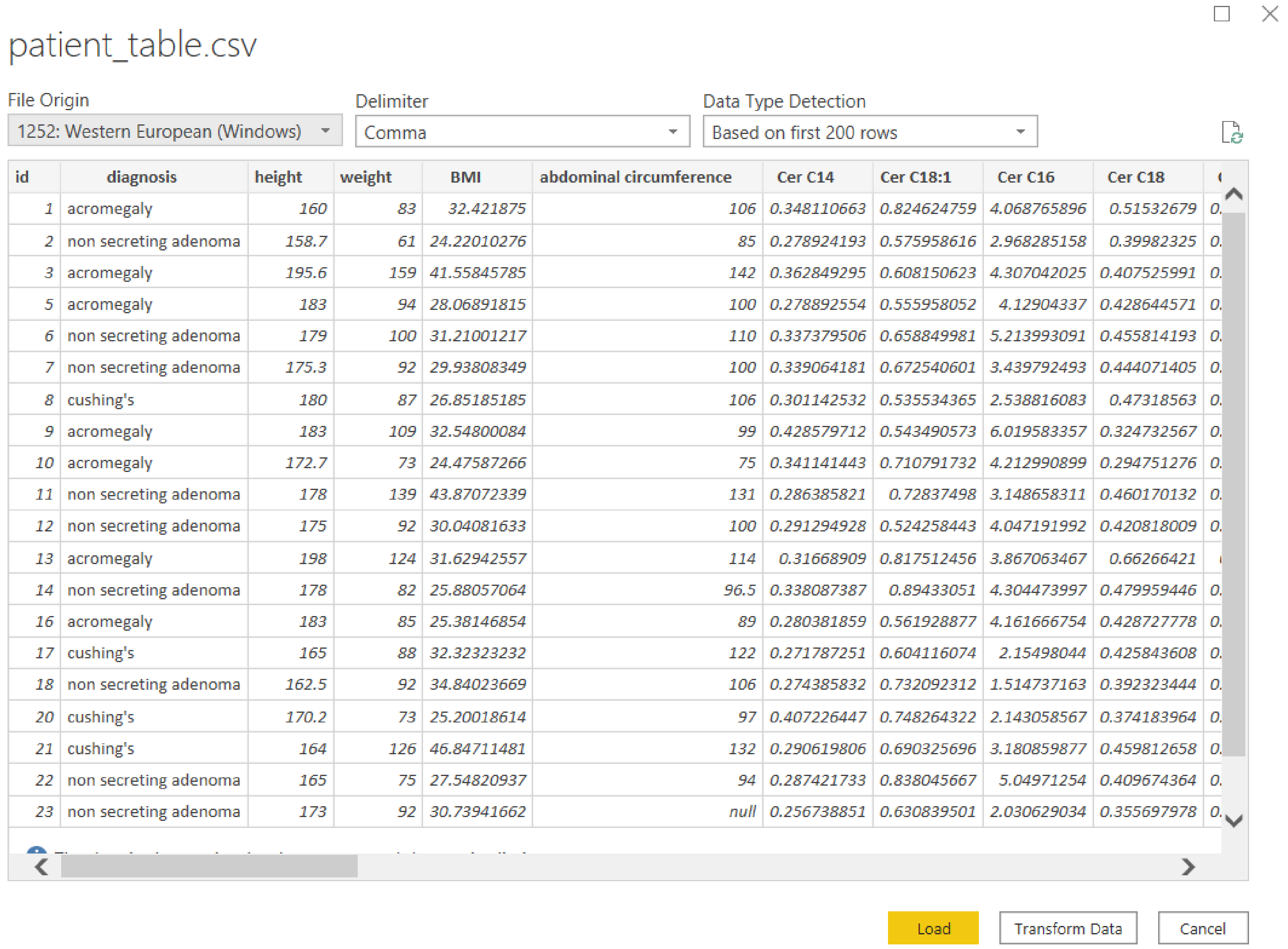
```
= Table.ReplaceValue(#"Filtered Rows1","non-functioning","Co
```

	patient_id	sample_id	group
1	1	12100	Acromegaly
2	2	12101	Control
3	3	12102	Acromegaly
4	5	12103	Acromegaly
5	6	12104	Control
6	7	12105	Control
7	9	12107	Acromegaly

4.5. Model View



5. Table 5: Patient Information Table



patient_table.csv

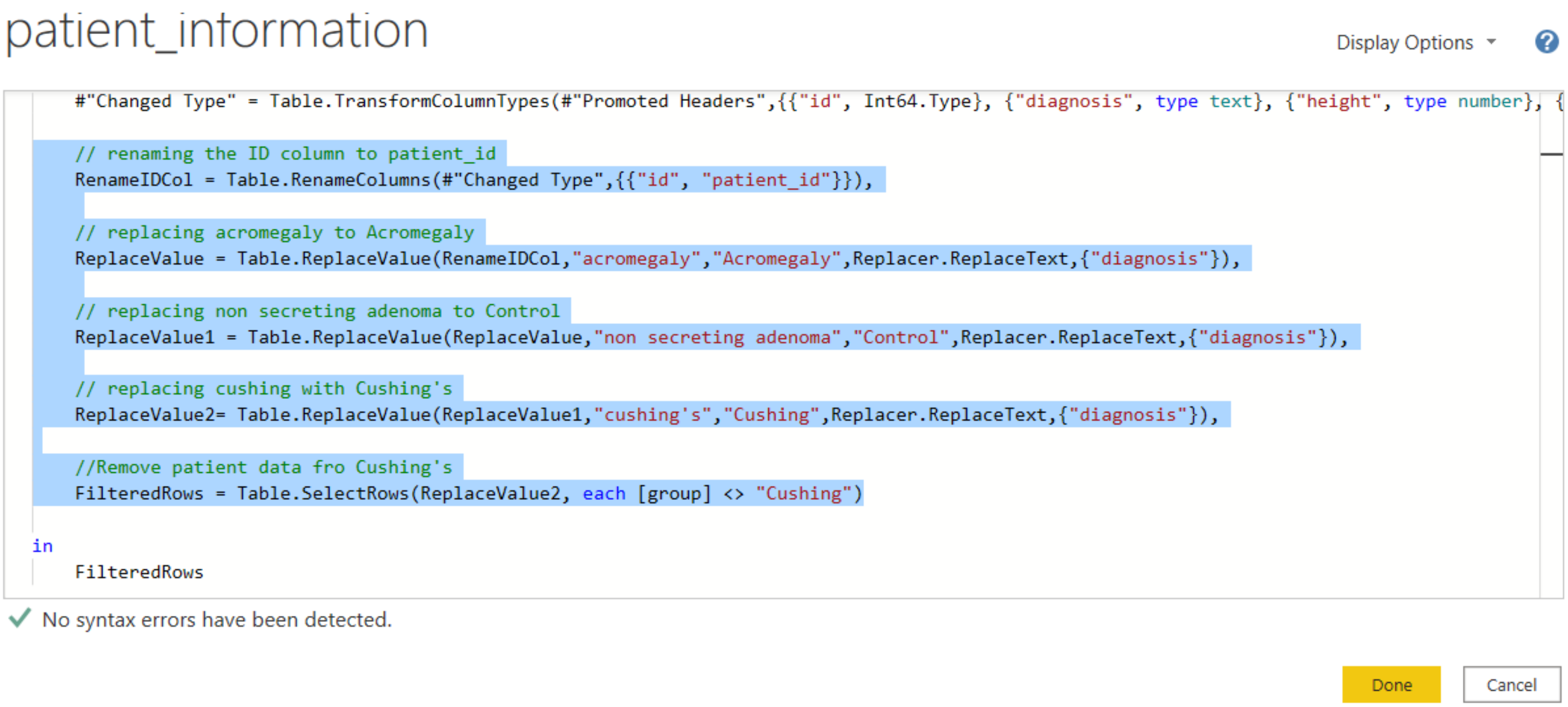
File Origin: 1252: Western European (Windows)

Delimiter: Comma

Data Type Detection: Based on first 200 rows

id	diagnosis	height	weight	BMI	abdominal circumference	Cer C14	Cer C18:1	Cer C16	Cer C18	(
1	acromegaly	160	83	32.421875	106	0.348110663	0.824624759	4.068765896	0.51532679	0.
2	non secreting adenoma	158.7	61	24.22010276	85	0.278924193	0.575958616	2.968285158	0.39982325	0.
3	acromegaly	195.6	159	41.55845785	142	0.362849295	0.608150623	4.307042025	0.407525991	0.
5	acromegaly	183	94	28.06891815	100	0.278892554	0.555958052	4.12904337	0.428644571	0.
6	non secreting adenoma	179	100	31.21001217	110	0.337379506	0.658849981	5.213993091	0.455814193	0.
7	non secreting adenoma	175.3	92	29.93808349	100	0.339064181	0.672540601	3.439792493	0.444071405	0.
8	cushing's	180	87	26.85185185	106	0.301142532	0.535534365	2.538816083	0.47318563	0.
9	acromegaly	183	109	32.54800084	99	0.428579712	0.543490573	6.019583357	0.324732567	0.
10	acromegaly	172.7	73	24.47587266	75	0.341141443	0.710791732	4.212990899	0.294751276	0.
11	non secreting adenoma	178	139	43.87072339	131	0.286385821	0.72837498	3.148658311	0.460170132	0.
12	non secreting adenoma	175	92	30.04081633	100	0.291294928	0.524258443	4.047191992	0.420818009	0.
13	acromegaly	198	124	31.62942557	114	0.31668909	0.817512456	3.867063467	0.66266421	(
14	non secreting adenoma	178	82	25.88057064	96.5	0.338087387	0.89433051	4.304473997	0.479959446	0.
16	acromegaly	183	85	25.38146854	89	0.280381859	0.561928877	4.161666754	0.428727778	0.
17	cushing's	165	88	32.32323232	122	0.271787251	0.604116074	2.15498044	0.425843608	0.
18	non secreting adenoma	162.5	92	34.84023669	106	0.274385832	0.732092312	1.514737163	0.392323444	0.
20	cushing's	170.2	73	25.20018614	97	0.407226447	0.748264322	2.143058567	0.374183964	0.
21	cushing's	164	126	46.84711481	132	0.290619806	0.690325696	3.180859877	0.459812658	0.
22	non secreting adenoma	165	75	27.54820937	94	0.287421733	0.838045667	5.04971254	0.409674364	0.
23	non secreting adenoma	173	92	30.73941662	null	0.256738851	0.630839501	2.030629034	0.355697978	0.

5.1. Replace Values

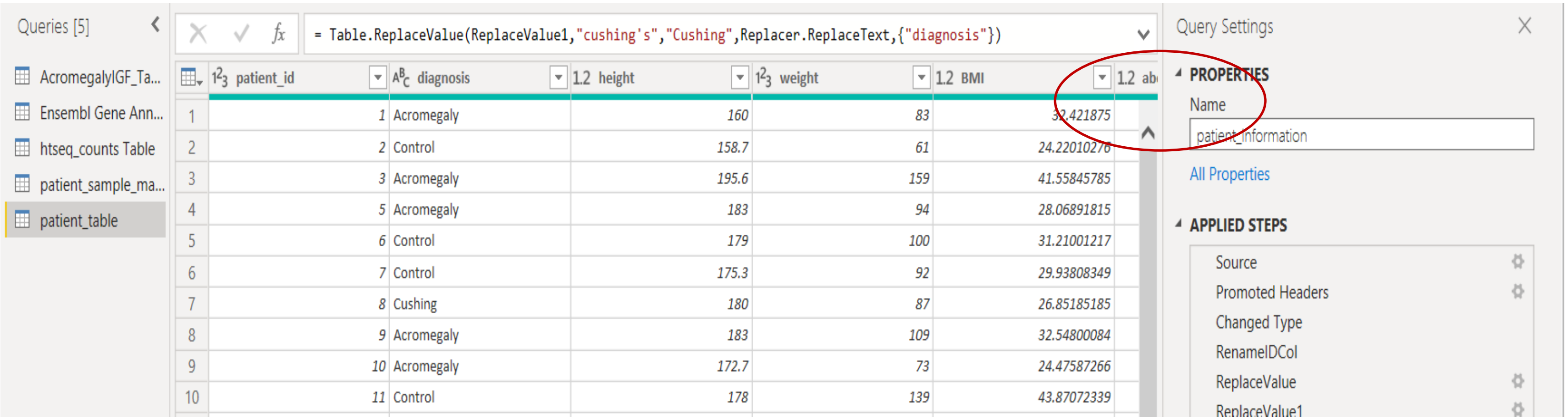


patient_information

```
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"id", Int64.Type}, {"diagnosis", type text}, {"height", type number}, {

    // renaming the ID column to patient_id
    RenameIDCol = Table.RenameColumns(#"Changed Type",{{"id", "patient_id"}}),

    // replacing acromegaly to Acromegaly
    ReplaceValue = Table.ReplaceValue(RenameIDCol,"acromegaly","Acromegaly",Replacer.ReplaceText,{"diagnosis"}),

    // replacing non secreting adenoma to Control
    ReplaceValue1 = Table.ReplaceValue(ReplaceValue,"non secreting adenoma","Control",Replacer.ReplaceText,{"diagnosis"}),

    // replacing cushing with Cushing's
    ReplaceValue2= Table.ReplaceValue(ReplaceValue1,"cushing's","Cushing",Replacer.ReplaceText,{"diagnosis"}),

    //Remove patient data fro Cushing's
    FilteredRows = Table.SelectRows(ReplaceValue2, each [group] <> "Cushing")

in
    FilteredRows
```

✓ No syntax errors have been detected.

5.2. Renaming the table



Queries [5]

- AcromegalyIGF_Ta...
- Ensembl Gene Ann...
- htseq_counts Table
- patient_sample_ma...
- patient_table

```
= Table.ReplaceValue(ReplaceValue1,"cushing's","Cushing",Replacer.ReplaceText,{"diagnosis"})
```

	patient_id	diagnosis	height	weight	BMI	ab
1	1	Acromegaly	160	83	32.421875	
2	2	Control	158.7	61	24.22010276	
3	3	Acromegaly	195.6	159	41.55845785	
4	5	Acromegaly	183	94	28.06891815	
5	6	Control	179	100	31.21001217	
6	7	Control	175.3	92	29.93808349	
7	8	Cushing	180	87	26.85185185	
8	9	Acromegaly	183	109	32.54800084	
9	10	Acromegaly	172.7	73	24.47587266	
10	11	Control	178	139	43.87072339	

Query Settings

PROPERTIES

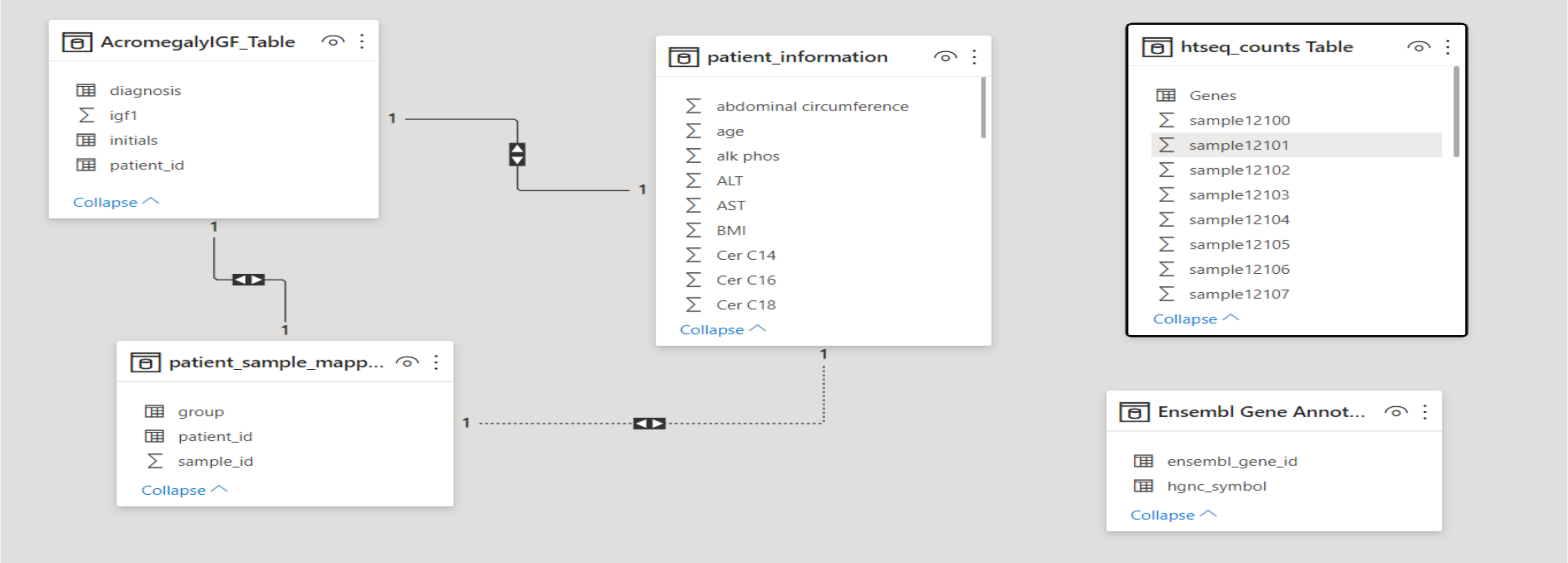
Name: patient_information

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- RenameIDCol
- ReplaceValue
- ReplaceValue1

5.3. Model View



AcromegalyIGF_Table
- diagnosis
- igf1
- initials
- patient_id

patient_information
- abdominal circumference
- age
- alk phos
- ALT
- AST
- BMI
- Cer C14
- Cer C16
- Cer C18

htseq_counts Table
- Genes
- sample12100
- sample12101
- sample12102
- sample12103
- sample12104
- sample12105
- sample12106
- sample12107

patient_sample_mapp...
- group
- patient_id
- sample_id

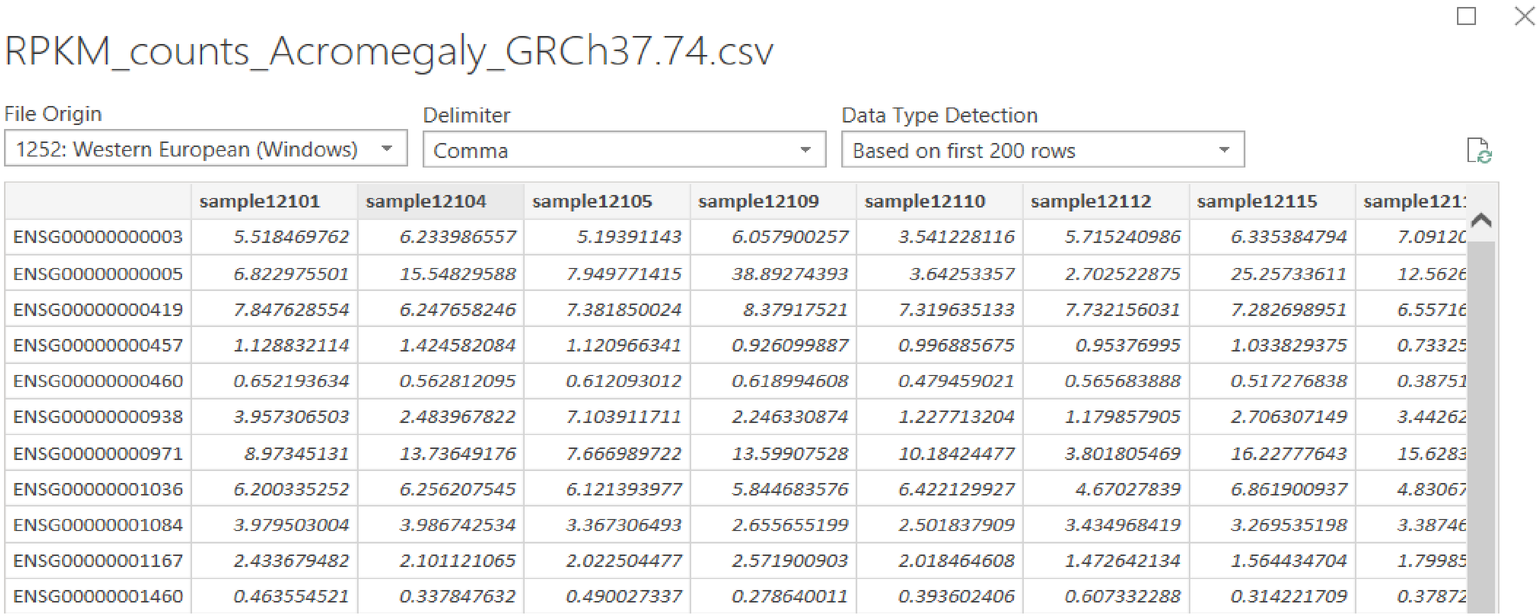
Ensembl Gene Annot...
- ensembl_gene_id
- hgnc_symbol

6. Table 6: IGF_RKPM_Count

6.1. Rename Column

Import the table as and rename the first column to “Genes”

RenameCols = Table.RenameColumns(#"Changed Type",{{"", "Genes"}})



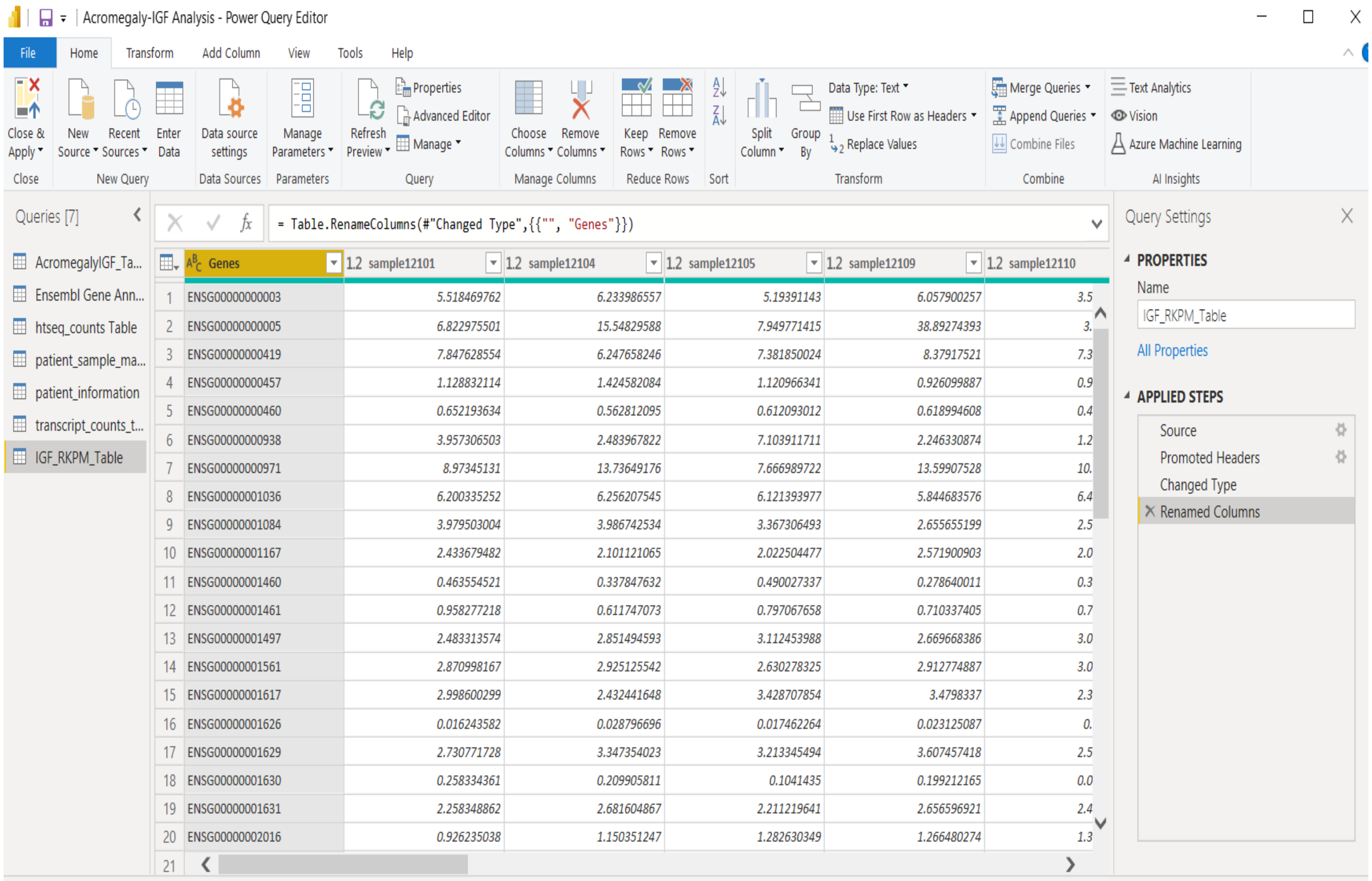
RPKM_counts_Acromegaly_GRCh37.74.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

Data Type Detection: Based on first 200 rows

	sample12101	sample12104	sample12105	sample12109	sample12110	sample12112	sample12115	sample121:
ENSG00000000003	5.518469762	6.233986557	5.19391143	6.057900257	3.541228116	5.715240986	6.335384794	7.0912C
ENSG00000000005	6.822975501	15.54829588	7.949771415	38.89274393	3.64253357	2.702522875	25.25733611	12.5626
ENSG00000000419	7.847628554	6.247658246	7.381850024	8.37917521	7.319635133	7.732156031	7.282698951	6.55716
ENSG00000000457	1.128832114	1.424582084	1.120966341	0.926099887	0.996885675	0.95376995	1.033829375	0.73325
ENSG00000000460	0.652193634	0.562812095	0.612093012	0.618994608	0.479459021	0.565683888	0.517276838	0.38751
ENSG00000000938	3.957306503	2.483967822	7.103911711	2.246330874	1.227713204	1.179857905	2.706307149	3.44262
ENSG00000000971	8.97345131	13.73649176	7.666989722	13.59907528	10.18424477	3.801805469	16.22777643	15.6283
ENSG00000001036	6.200335252	6.256207545	6.121393977	5.844683576	6.422129927	4.67027839	6.861900937	4.83067
ENSG00000001084	3.979503004	3.986742534	3.367306493	2.655655199	2.501837909	3.434968419	3.269535198	3.38746
ENSG00000001167	2.433679482	2.101121065	2.022504477	2.571900903	2.018464608	1.472642134	1.564434704	1.79985
ENSG00000001460	0.463554521	0.337847632	0.490027337	0.278640011	0.393602406	0.607332288	0.314221709	0.37872



Queries [7]

- AcromegalyIGF_Ta...
- Ensembl Gene Ann...
- htseq_counts Table
- patient_sample_ma...
- patient_information
- transcript_counts_t...
- IGF_RKPM_Table

```
= Table.RenameColumns(#"Changed Type",{{"", "Genes"}})
```

	Genes	sample12101	sample12104	sample12105	sample12109	sample12110
1	ENSG00000000003	5.518469762	6.233986557	5.19391143	6.057900257	3.5
2	ENSG00000000005	6.822975501	15.54829588	7.949771415	38.89274393	3.
3	ENSG00000000419	7.847628554	6.247658246	7.381850024	8.37917521	7.3
4	ENSG00000000457	1.128832114	1.424582084	1.120966341	0.926099887	0.9
5	ENSG00000000460	0.652193634	0.562812095	0.612093012	0.618994608	0.4
6	ENSG00000000938	3.957306503	2.483967822	7.103911711	2.246330874	1.2
7	ENSG00000000971	8.97345131	13.73649176	7.666989722	13.59907528	10.
8	ENSG00000001036	6.200335252	6.256207545	6.121393977	5.844683576	6.4
9	ENSG00000001084	3.979503004	3.986742534	3.367306493	2.655655199	2.5
10	ENSG00000001167	2.433679482	2.101121065	2.022504477	2.571900903	2.0
11	ENSG00000001460	0.463554521	0.337847632	0.490027337	0.278640011	0.3
12	ENSG00000001461	0.958277218	0.611747073	0.797067658	0.710337405	0.7
13	ENSG00000001497	2.483313574	2.851494593	3.112453988	2.669668386	3.0
14	ENSG00000001561	2.870998167	2.925125542	2.630278325	2.912774887	3.0
15	ENSG00000001617	2.998600299	2.432441648	3.428707854	3.4798337	2.3
16	ENSG00000001626	0.016243582	0.028796696	0.017462264	0.023125087	0.
17	ENSG00000001629	2.730771728	3.347354023	3.213345494	3.607457418	2.5
18	ENSG00000001630	0.258334361	0.209905811	0.1041435	0.199212165	0.0
19	ENSG00000001631	2.258348862	2.681604867	2.211219641	2.656596921	2.4
20	ENSG00000002016	0.926235038	1.150351247	1.282630349	1.266480274	1.3
21						

Query Settings

PROPERTIES

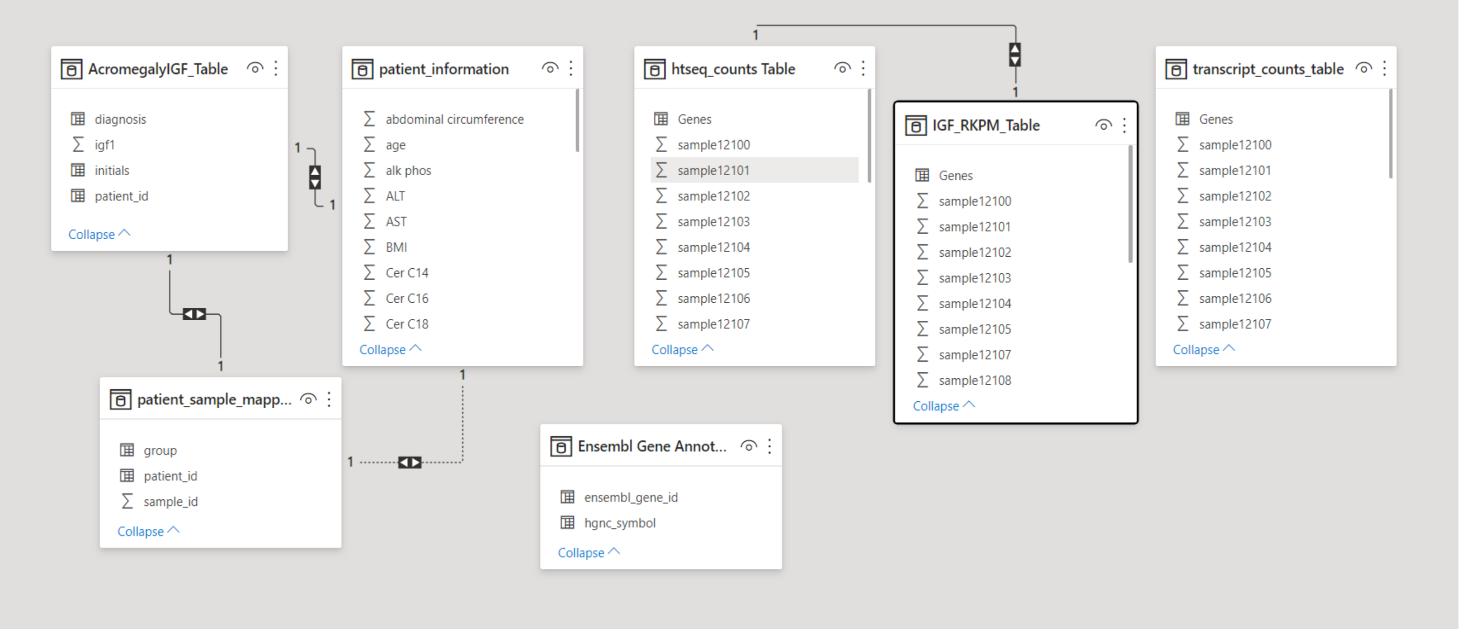
Name: IGF_RKPM_Table

All Properties

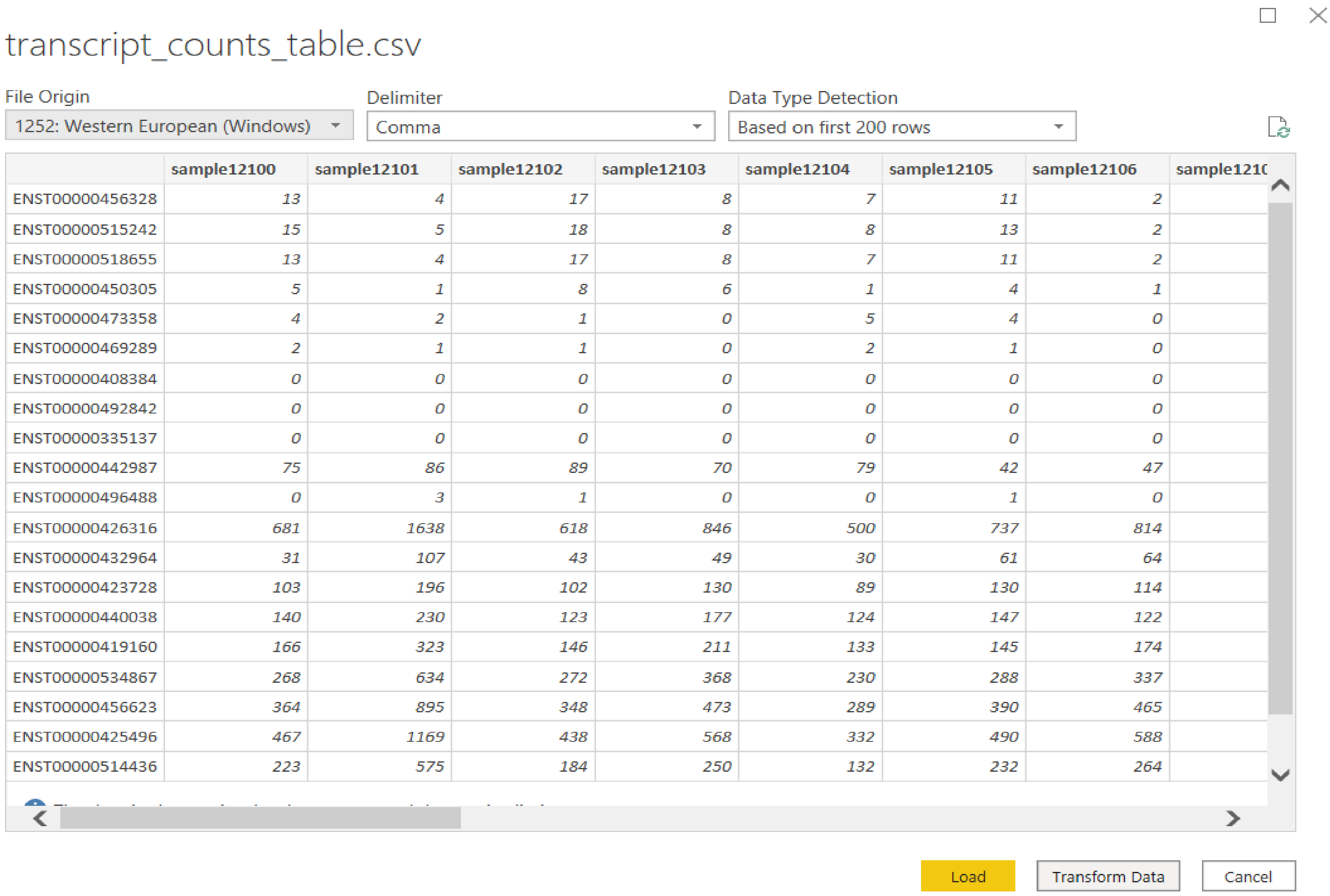
APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Renamed Columns

6.2. Model View



7. Table 7: Transcripts Count



transcript_counts_table.csv

File Origin: 1252: Western European (Windows)

Delimiter: Comma

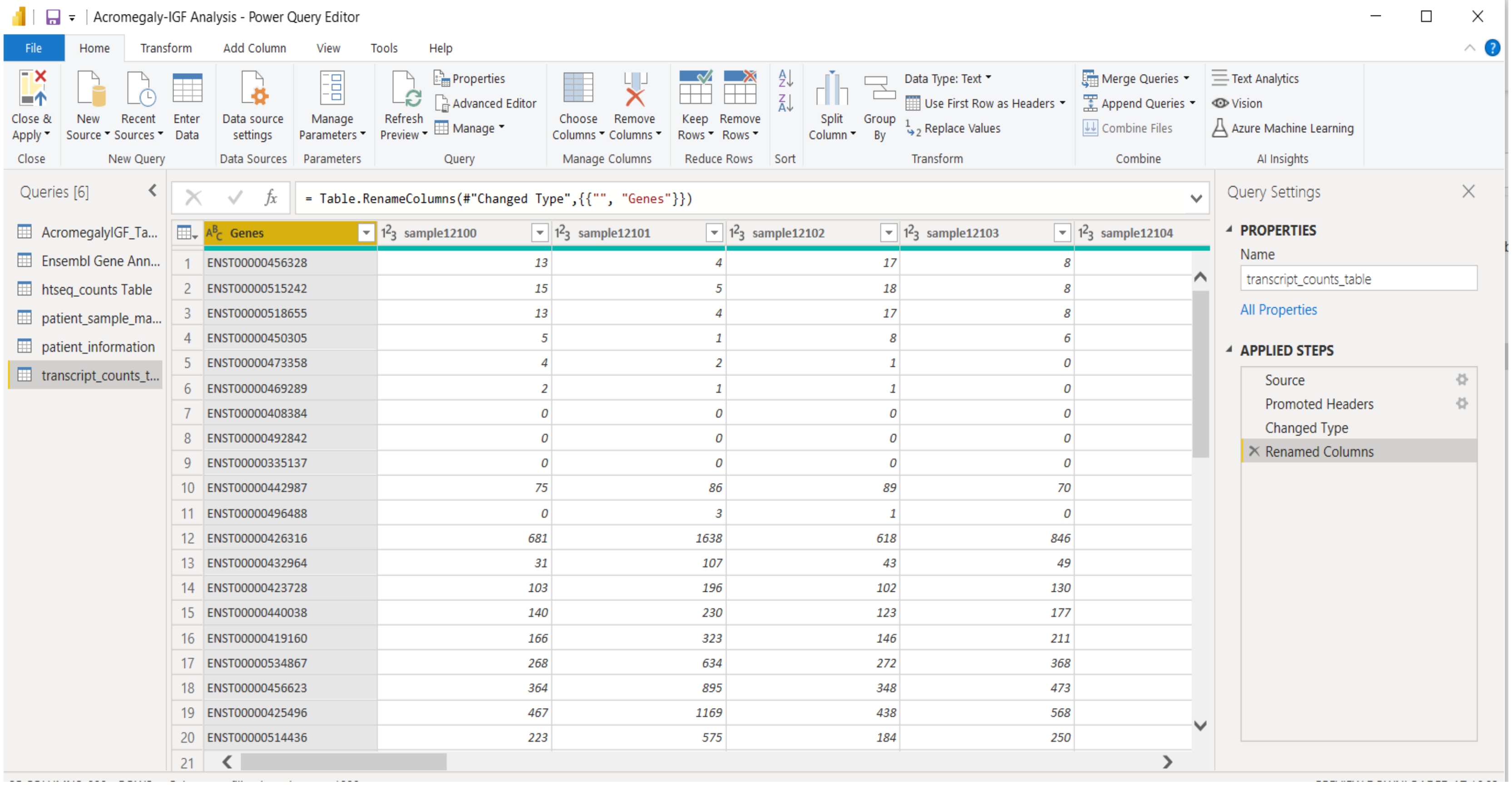
Data Type Detection: Based on first 200 rows

	sample12100	sample12101	sample12102	sample12103	sample12104	sample12105	sample12106	sample1210
ENST00000456328	13	4	17	8	7	11	2	
ENST00000515242	15	5	18	8	8	13	2	
ENST00000518655	13	4	17	8	7	11	2	
ENST00000450305	5	1	8	6	1	4	1	
ENST00000473358	4	2	1	0	5	4	0	
ENST00000469289	2	1	1	0	2	1	0	
ENST00000408384	0	0	0	0	0	0	0	
ENST00000492842	0	0	0	0	0	0	0	
ENST00000335137	0	0	0	0	0	0	0	
ENST00000442987	75	86	89	70	79	42	47	
ENST00000496488	0	3	1	0	0	1	0	
ENST00000426316	681	1638	618	846	500	737	814	
ENST00000432964	31	107	43	49	30	61	64	
ENST00000423728	103	196	102	130	89	130	114	
ENST00000440038	140	230	123	177	124	147	122	
ENST00000419160	166	323	146	211	133	145	174	
ENST00000534867	268	634	272	368	230	288	337	
ENST00000456623	364	895	348	473	289	390	465	
ENST00000425496	467	1169	438	568	332	490	588	
ENST00000514436	223	575	184	250	132	232	264	

7.1. Rename Column

Import the table as and rename the first column to “Genes”

RenameCols = Table.RenameColumns(#"Changed Type",{{"", "Genes"}})



```
= Table.RenameColumns(#"Changed Type",{{"", "Genes"}})
```

	Genes	sample12100	sample12101	sample12102	sample12103	sample12104
1	ENST00000456328	13	4	17	8	
2	ENST00000515242	15	5	18	8	
3	ENST00000518655	13	4	17	8	
4	ENST00000450305	5	1	8	6	
5	ENST00000473358	4	2	1	0	
6	ENST00000469289	2	1	1	0	
7	ENST00000408384	0	0	0	0	
8	ENST00000492842	0	0	0	0	
9	ENST00000335137	0	0	0	0	
10	ENST00000442987	75	86	89	70	
11	ENST00000496488	0	3	1	0	
12	ENST00000426316	681	1638	618	846	
13	ENST00000432964	31	107	43	49	
14	ENST00000423728	103	196	102	130	
15	ENST00000440038	140	230	123	177	
16	ENST00000419160	166	323	146	211	
17	ENST00000534867	268	634	272	368	
18	ENST00000456623	364	895	348	473	
19	ENST00000425496	467	1169	438	568	
20	ENST00000514436	223	575	184	250	
21						

PROPERTIES

Name: transcript_counts_table

All Properties

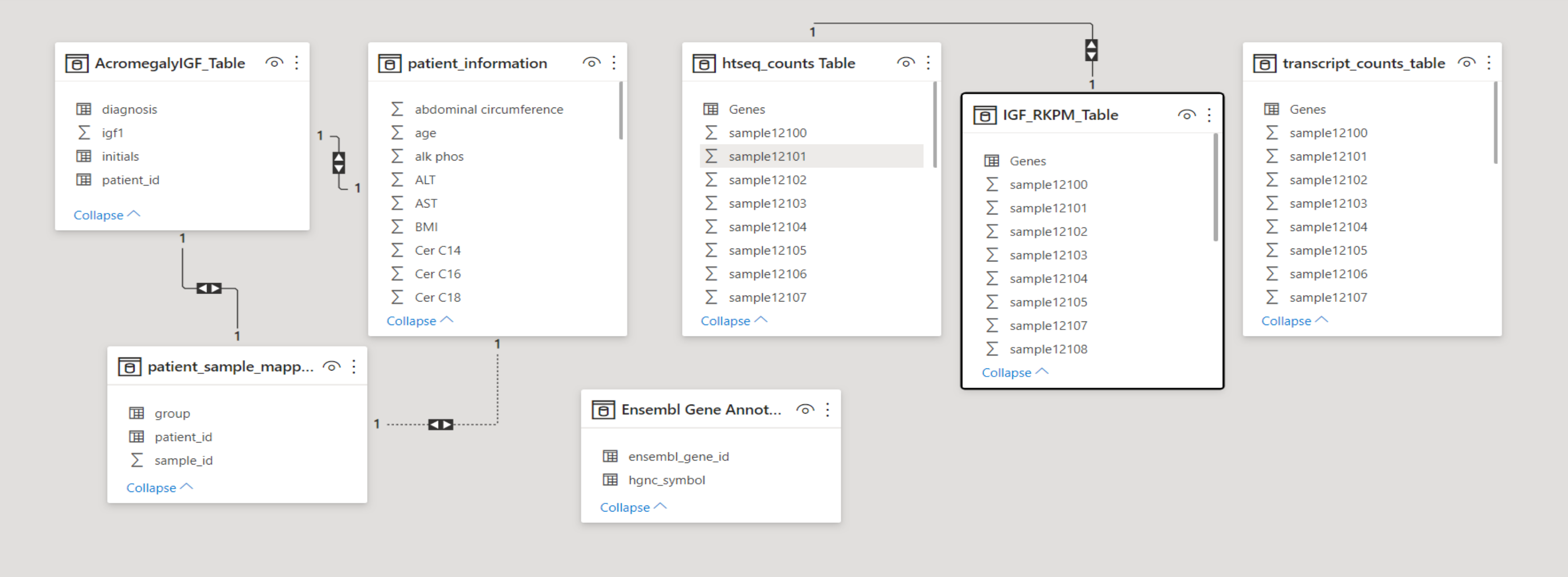
APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Renamed Columns

C) Data Modelling –Schema Facts and Dimensions

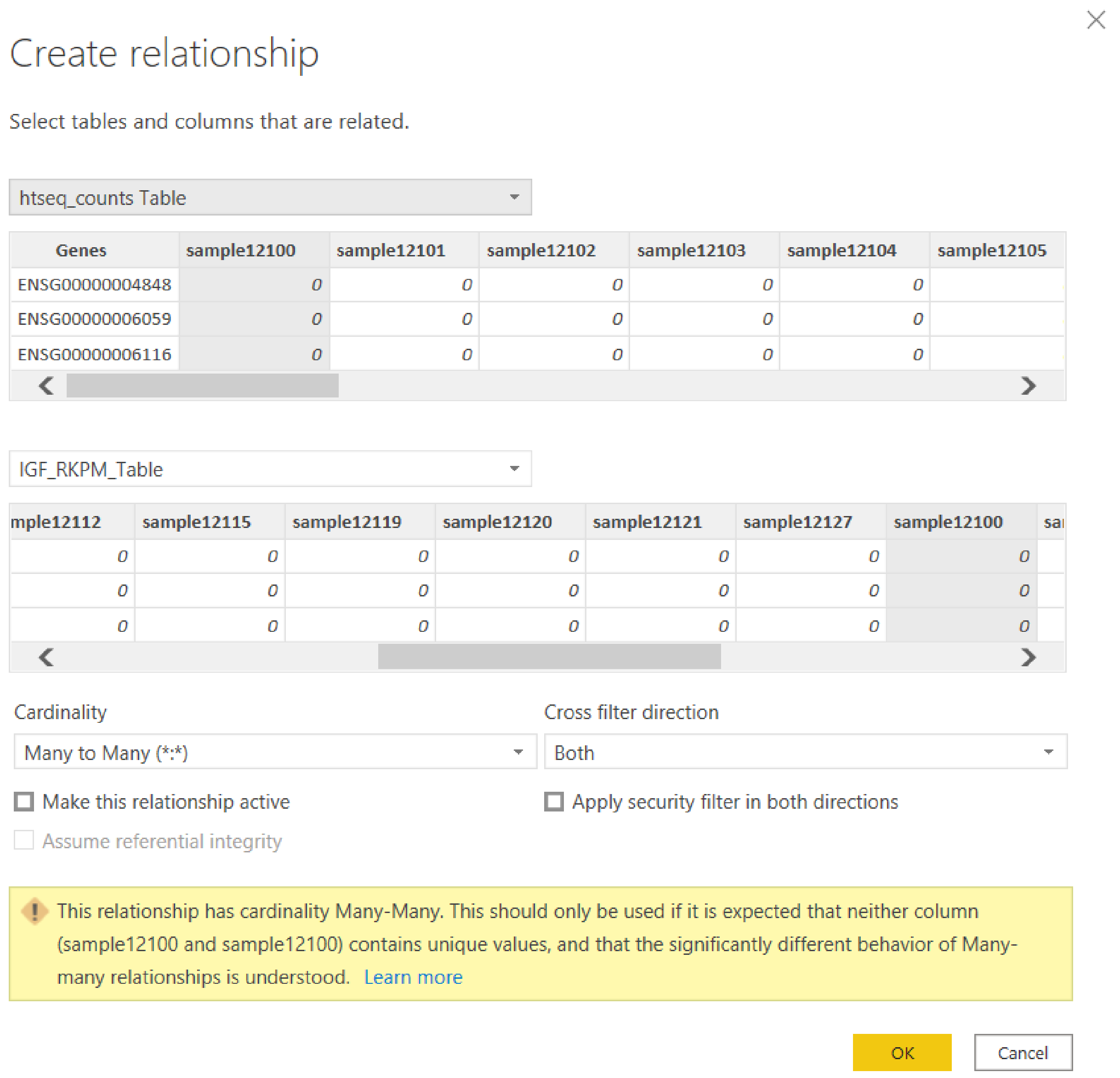
1. Data Modelling Process

The data model after loading the tables looks as below. On further investigating the relationships shown in the model, the tables are not connected correctly.



1.1. Creating Relationships

When we try to map the sample ID from the “**Patient_sample_mapping**” table and **htseq_counts_Table, IGF_RKPM_Table, transcript_counts_table**. The model relationships are mapped as below. When trying to map the sample from htseq_counts_Table and IGF_RKPM_Table, this leads many to many relationships.



Create relationship

Select tables and columns that are related.

htseq_counts Table

Genes	sample12100	sample12101	sample12102	sample12103	sample12104	sample12105
ENSG00000004848	0	0	0	0	0	
ENSG00000006059	0	0	0	0	0	
ENSG00000006116	0	0	0	0	0	

IGF_RKPM_Table

mple12112	sample12115	sample12119	sample12120	sample12121	sample12127	sample12100	sa
0	0	0	0	0	0	0	
0	0	0	0	0	0	0	
0	0	0	0	0	0	0	

Cardinality: Many to Many (*:*)

Cross filter direction: Both

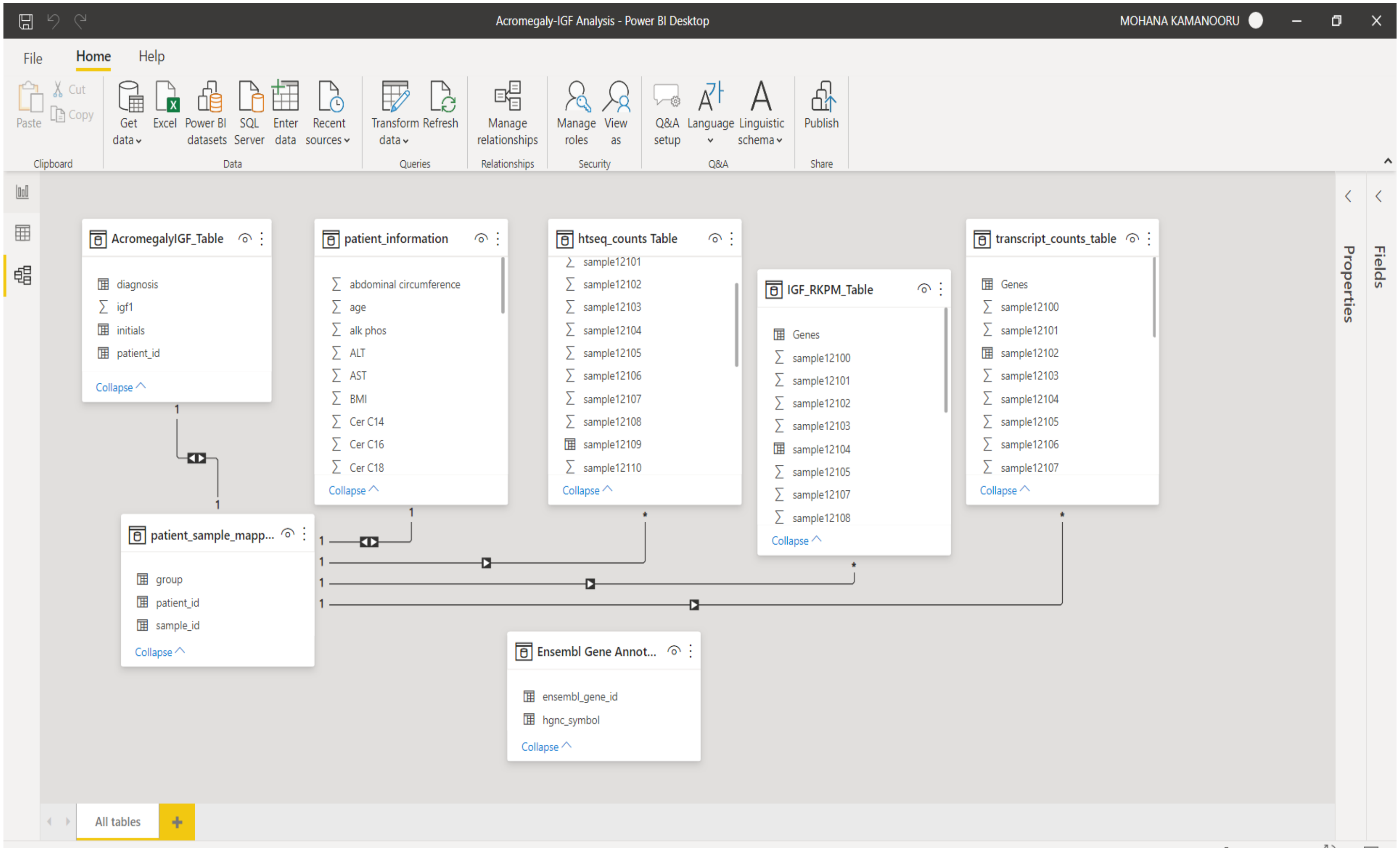
□ Make this relationship active

□ Apply security filter in both directions

□ Assume referential integrity

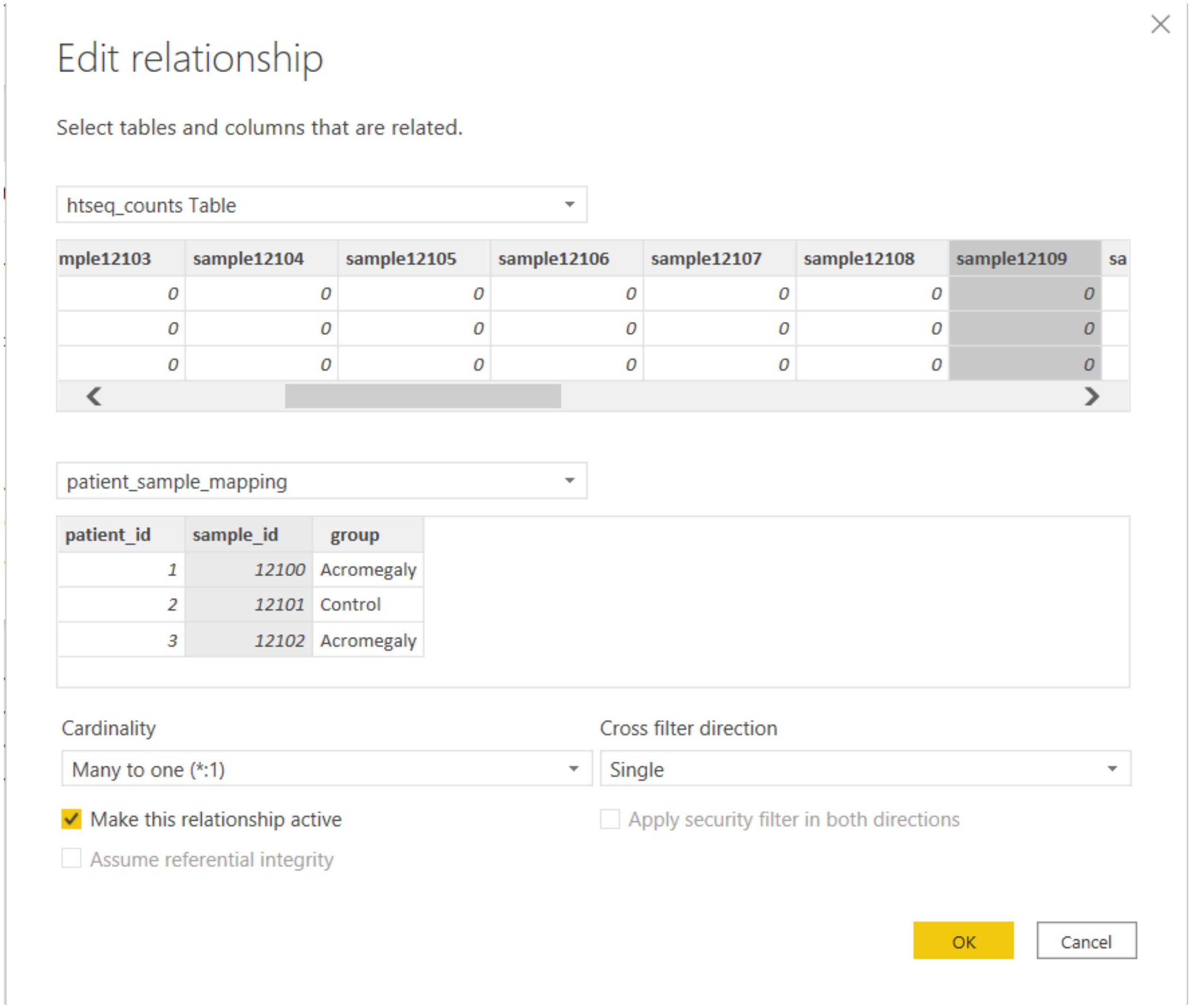
This relationship has cardinality Many-Many. This should only be used if it is expected that neither column (sample12100 and sample12100) contains unique values, and that the significantly different behavior of Many-many relationships is understood. Learn more

To avoid this, we transform our data into tables a bit more as shown below.



1.2. Edit Relationships

Right-click on the relationship and select edit properties, we see the image below. Where the relationships are incorrectly mapped. Sample12109 is mapped with sample_id. But the correct relationship is sample_id column should be mapped to the column names in htseq_counts Table. This can be achieved by unpivoting the table.



Edit relationship

Select tables and columns that are related.

htseq_counts Table

mple12103	sample12104	sample12105	sample12106	sample12107	sample12108	sample12109	sa
0	0	0	0	0	0	0	
0	0	0	0	0	0	0	
0	0	0	0	0	0	0	

patient_sample_mapping

patient_id	sample_id	group
1	12100	Acromegaly
2	12101	Control
3	12102	Acromegaly

Cardinality: Many to one (*:1)

Cross filter direction: Single

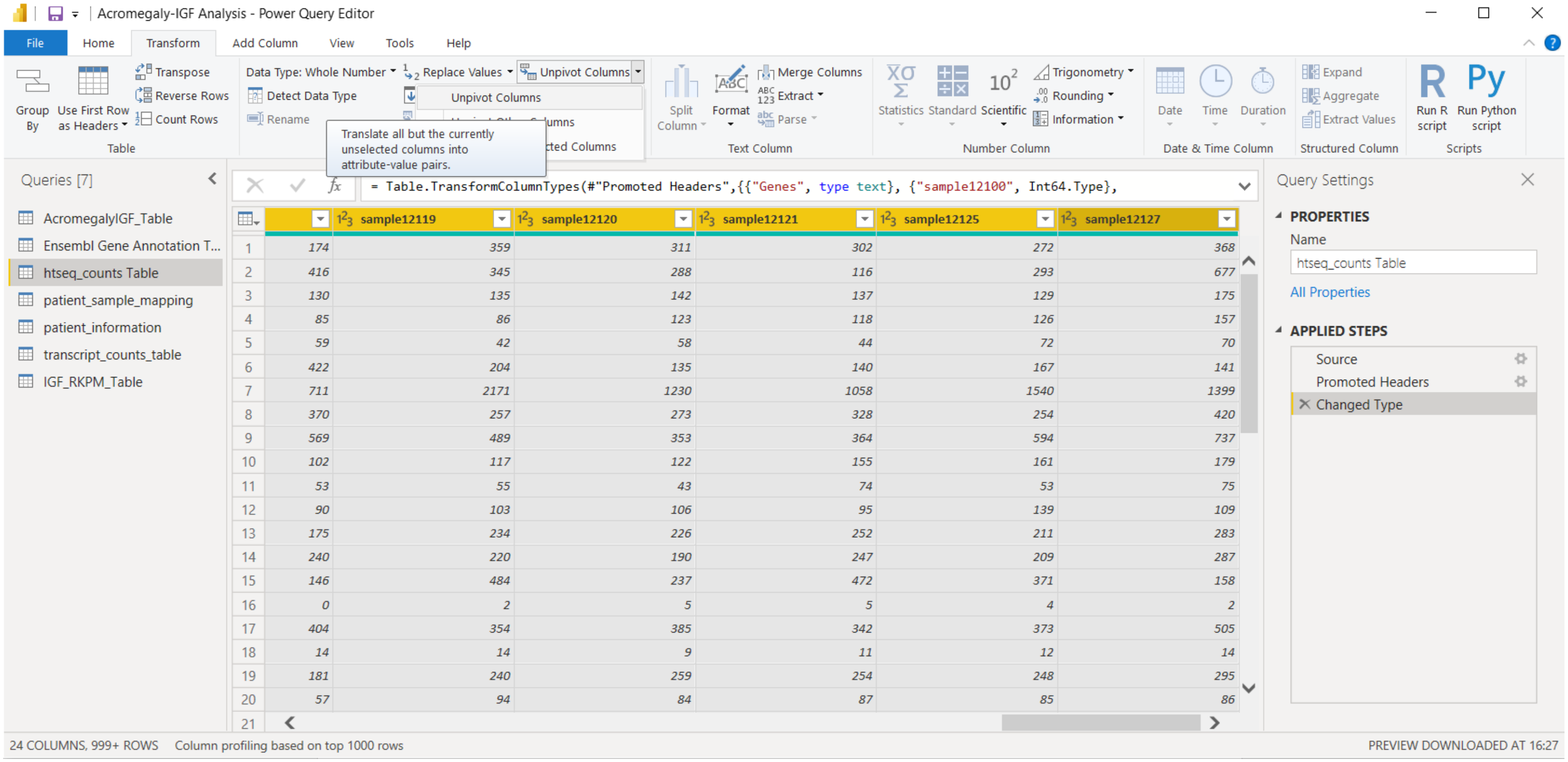
☑ Make this relationship active

□ Apply security filter in both directions

□ Assume referential integrity

1.3. Unpivoting Columns

Select all the sample columns from the **htseq_counts Table** and click on **Unpivot Columns** as shown below.



Translate all but the currently unselected columns into attribute-value pairs.

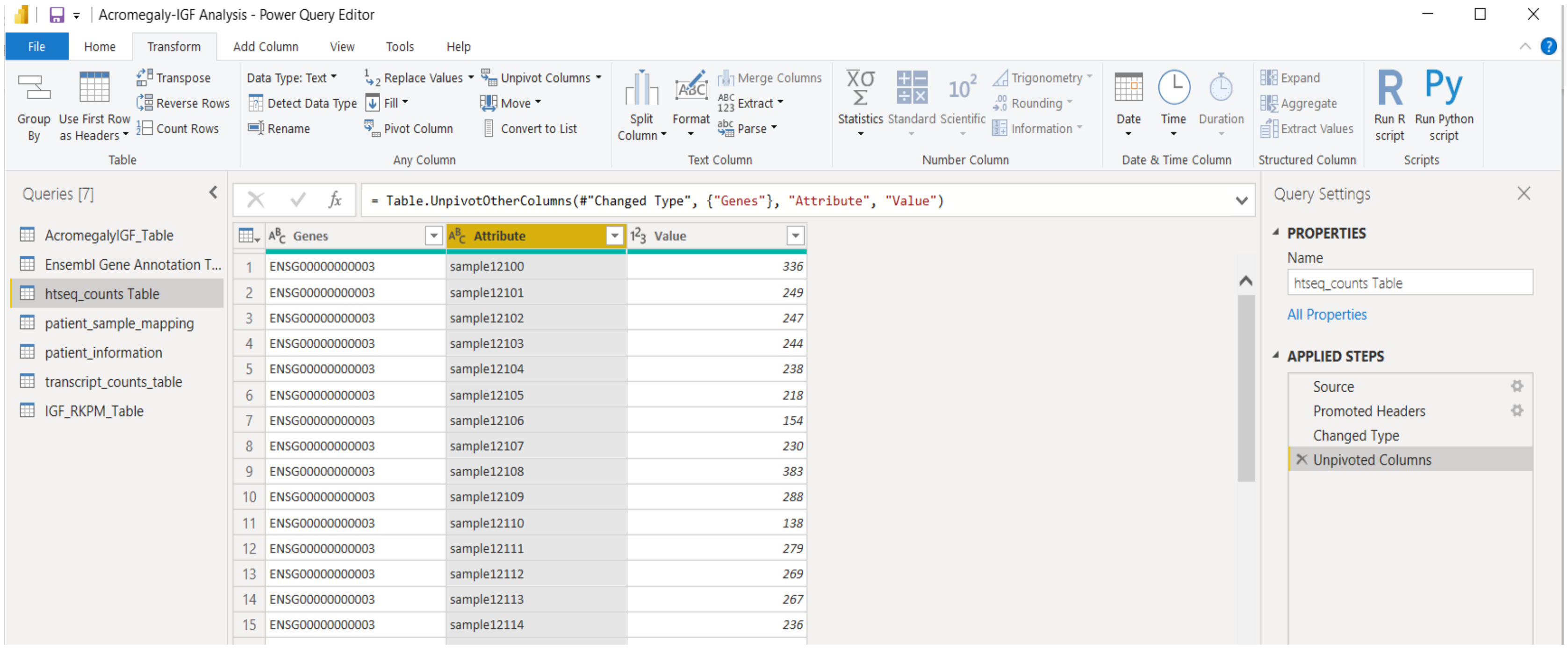
```
= Table.TransformColumnTypes(#"Promoted Headers",{{"Genes", type text}, {"sample12100", Int64.Type},
```

Name: htseq_counts Table

		sample12119	sample12120	sample12121	sample12125	sample12127
1	174	359	311	302	272	368
2	416	345	288	116	293	677
3	130	135	142	137	129	175
4	85	86	123	118	126	157
5	59	42	58	44	72	70
6	422	204	135	140	167	141
7	711	2171	1230	1058	1540	1399
8	370	257	273	328	254	420
9	569	489	353	364	594	737
10	102	117	122	155	161	179
11	53	55	43	74	53	75
12	90	103	106	95	139	109
13	175	234	226	252	211	283
14	240	220	190	247	209	287
15	146	484	237	472	371	158
16	0	2	5	5	4	2
17	404	354	385	342	373	505
18	14	14	9	11	12	14
19	181	240	259	254	248	295
20	57	94	84	87	85	86
21						

24 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED AT 16:27

Table transforms as below, now rename the **Attribute** and **Value** Columns to **sample** and **counts** respectively.



Queries [7]

- AcromegalyIGF_Table
- Ensembl Gene Annotation T...
- htseq_counts Table
- patient_sample_mapping
- patient_information
- transcript_counts_table
- IGF_RKPM_Table

```
= Table.UnpivotOtherColumns(#"Changed Type", {"Genes"}, "Attribute", "Value")
```

	Genes	Attribute	Value
1	ENSG00000000003	sample12100	336
2	ENSG00000000003	sample12101	249
3	ENSG00000000003	sample12102	247
4	ENSG00000000003	sample12103	244
5	ENSG00000000003	sample12104	238
6	ENSG00000000003	sample12105	218
7	ENSG00000000003	sample12106	154
8	ENSG00000000003	sample12107	230
9	ENSG00000000003	sample12108	383
10	ENSG00000000003	sample12109	288
11	ENSG00000000003	sample12110	138
12	ENSG00000000003	sample12111	279
13	ENSG00000000003	sample12112	269
14	ENSG00000000003	sample12113	267
15	ENSG00000000003	sample12114	236

Name: htseq_counts Table

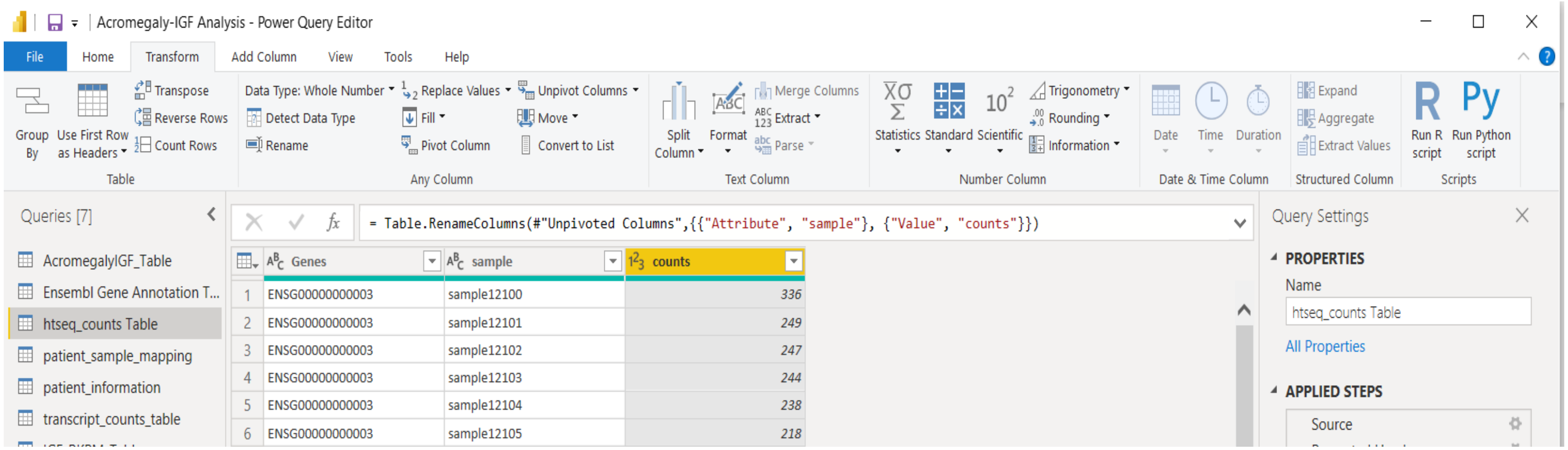
APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Unpivoted Columns

1.4. Renaming Columns

Renaming Attribute and Value Column using M formula,

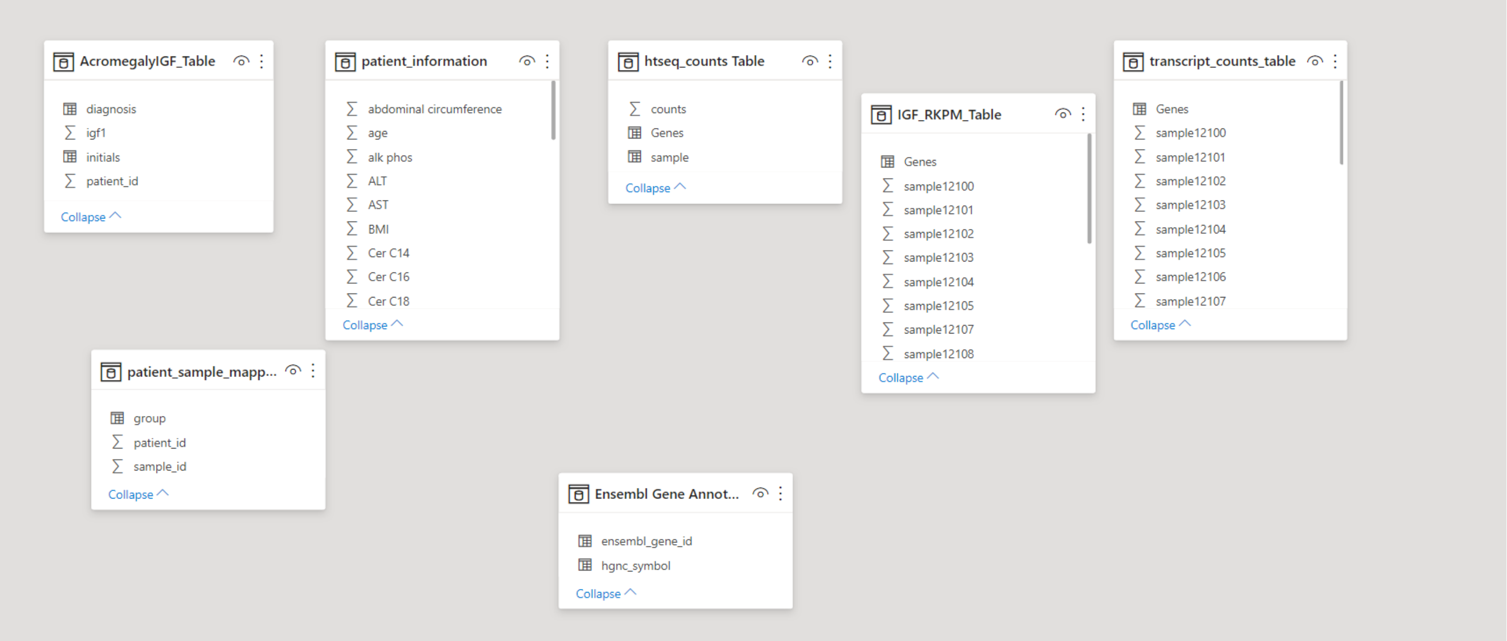
= Table.RenameColumns(#"Unpivoted Columns",{{"Attribute", "sample"}, {"Value", "counts"}})



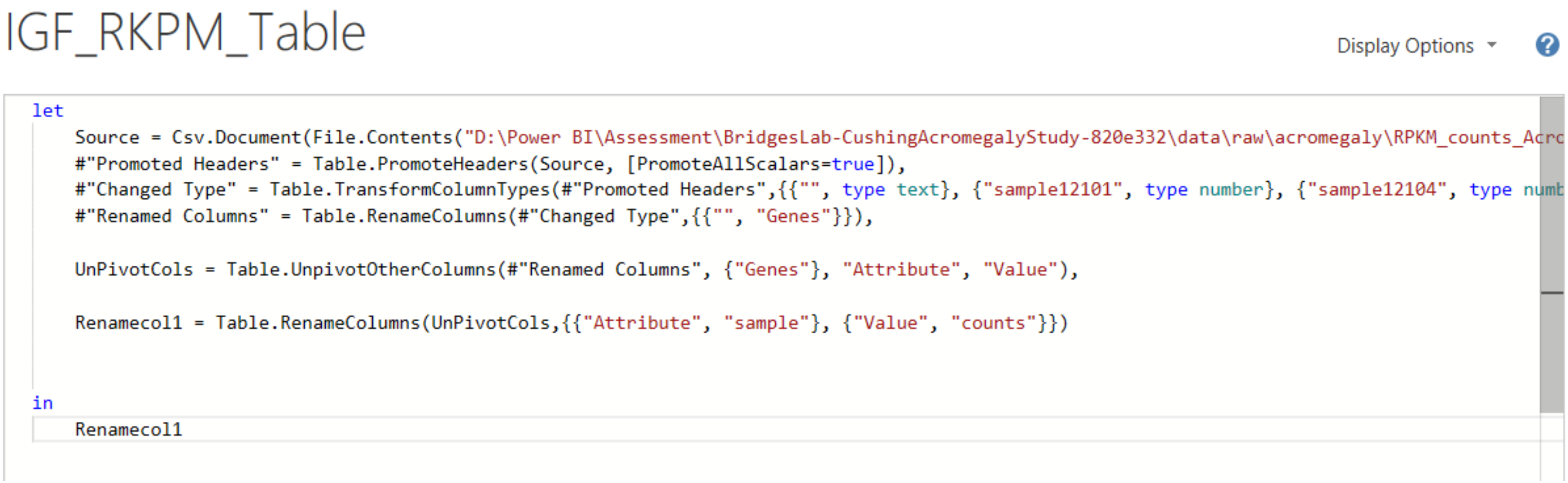
```
= Table.RenameColumns(#"Unpivoted Columns",{{"Attribute", "sample"}, {"Value", "counts"}})
```

	Genes	sample	counts
1	ENSG00000000003	sample12100	336
2	ENSG00000000003	sample12101	249
3	ENSG00000000003	sample12102	247
4	ENSG00000000003	sample12103	244
5	ENSG00000000003	sample12104	238
6	ENSG00000000003	sample12105	218

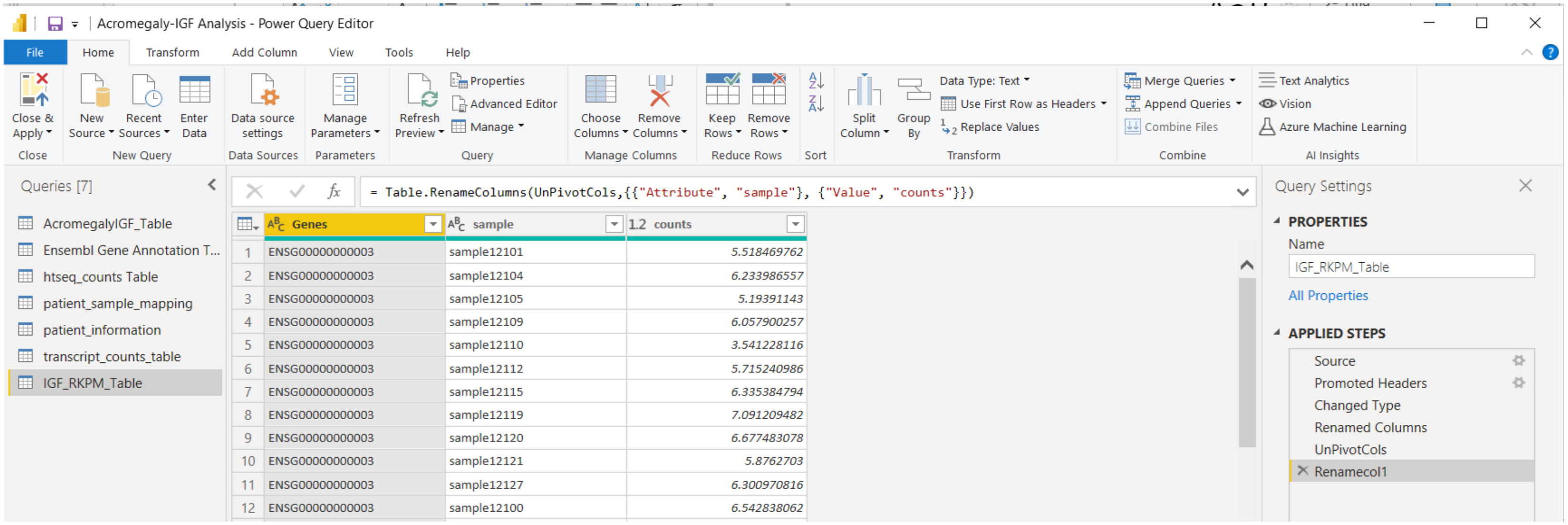
The tables look as below after unpivoting the columns in htseq_counts Table.



Repeating the unpivot and rename column steps in **IGF_RKPM_Table**.



```
let
    Source = Csv.Document(File.Contents("D:\Power BI\Assessment\BridgesLab-CushingAcromegalyStudy-820e332\data\raw\acromegaly\RPKM_counts_Acro
    #"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"", type text}, {"sample12101", type number}, {"sample12104", type numb
    #"Renamed Columns" = Table.RenameColumns(#"Changed Type",{{"", "Genes"}}),

    UnPivotCols = Table.UnpivotOtherColumns(#"Renamed Columns", {"Genes"}, "Attribute", "Value"),

    Renamecol1 = Table.RenameColumns(UnPivotCols,{{"Attribute", "sample"}, {"Value", "counts"}})


in
    Renamecol1
```



Queries [7]

- AcromegalyIGF_Table
- Ensembl Gene Annotation T...
- htseq_counts Table
- patient_sample_mapping
- patient_information
- transcript_counts_table
- IGF_RKPM_Table

```
= Table.RenameColumns(UnPivotCols,{{"Attribute", "sample"}, {"Value", "counts"}})
```

	Genes	sample	counts
1	ENSG00000000003	sample12101	5.518469762
2	ENSG00000000003	sample12104	6.233986557
3	ENSG00000000003	sample12105	5.19391143
4	ENSG00000000003	sample12109	6.057900257
5	ENSG00000000003	sample12110	3.541228116
6	ENSG00000000003	sample12112	5.715240986
7	ENSG00000000003	sample12115	6.335384794
8	ENSG00000000003	sample12119	7.091209482
9	ENSG00000000003	sample12120	6.677483078
10	ENSG00000000003	sample12121	5.8762703
11	ENSG00000000003	sample12127	6.300970816
12	ENSG00000000003	sample12100	6.542838062

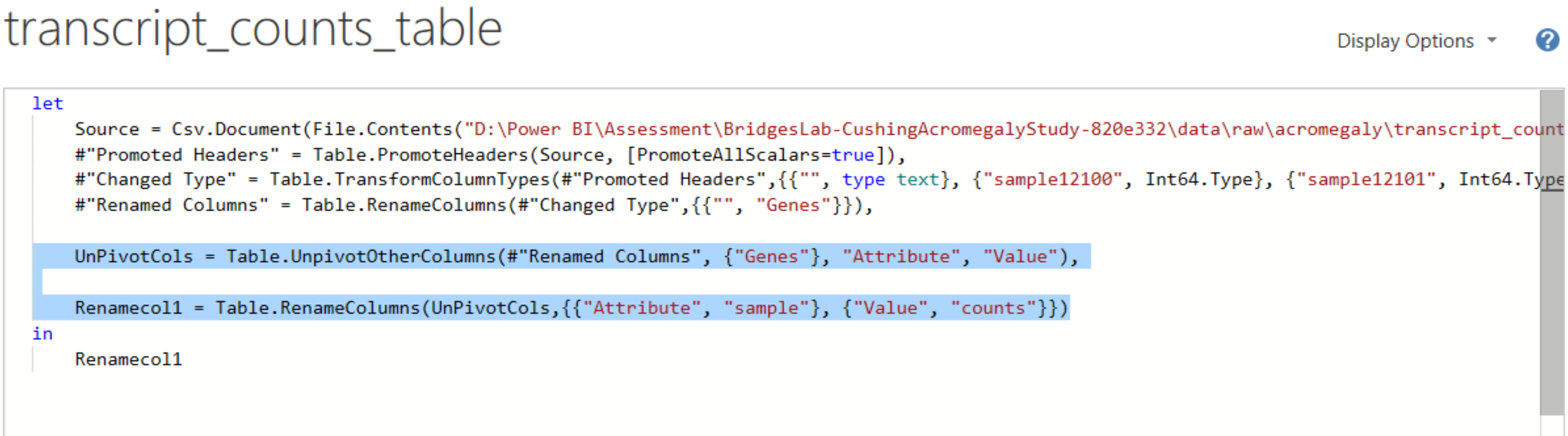
PROPERTIES

Name: IGF_RKPM_Table

APPLIED STEPS

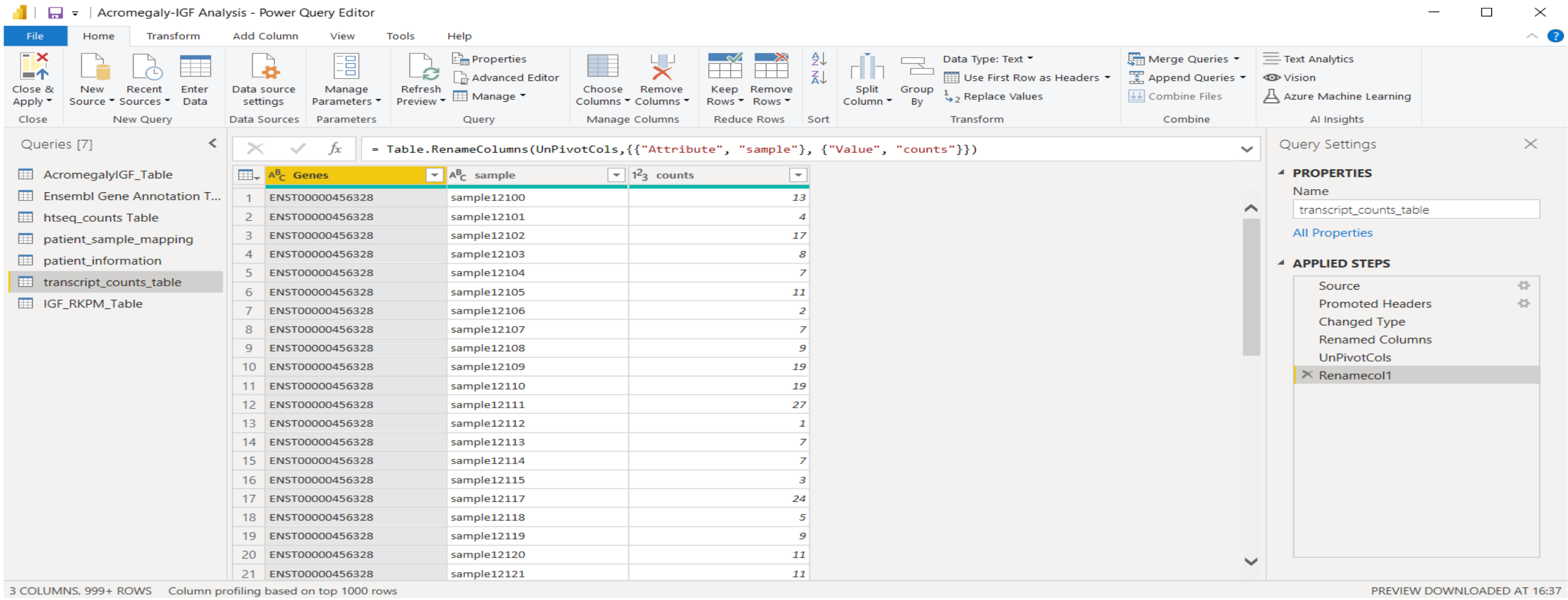
- Source
- Promoted Headers
- Changed Type
- Renamed Columns
- UnPivotCols
- Renamecol1

Repeating the unpivot and rename column steps in **transcript_counts_Table**.



transcript_counts_table

```
let
    Source = Csv.Document(File.Contents("D:\Power BI\Assessment\BridgesLab-CushingAcromegalyStudy-820e332\data\raw\acromegaly\transcript_count
    #"Promoted Headers" = Table.PromoteHeaders(Source, [PromoteAllScalars=true]),
    #"Changed Type" = Table.TransformColumnTypes(#"Promoted Headers",{{"", type text}, {"sample12100", Int64.Type}, {"sample12101", Int64.Type
    #"Renamed Columns" = Table.RenameColumns(#"Changed Type",{{"", "Genes"}}),

    UnPivotCols = Table.UnpivotOtherColumns(#"Renamed Columns", {"Genes"}, "Attribute", "Value"),

    Renamecol1 = Table.RenameColumns(UnPivotCols,{{"Attribute", "sample"}, {"Value", "counts"}})
in
    Renamecol1
```



Queries [7]

- AcromegalyIGF_Table
- Ensembl Gene Annotation T...
- htseq_counts Table
- patient_sample_mapping
- patient_information
- transcript_counts_table
- IGF_RKPM_Table

```
= Table.RenameColumns(UnPivotCols,{{"Attribute", "sample"}, {"Value", "counts"}})
```

	Genes	sample	counts
1	ENST00000456328	sample12100	13
2	ENST00000456328	sample12101	4
3	ENST00000456328	sample12102	17
4	ENST00000456328	sample12103	8
5	ENST00000456328	sample12104	7
6	ENST00000456328	sample12105	11
7	ENST00000456328	sample12106	2
8	ENST00000456328	sample12107	7
9	ENST00000456328	sample12108	9
10	ENST00000456328	sample12109	19
11	ENST00000456328	sample12110	19
12	ENST00000456328	sample12111	27
13	ENST00000456328	sample12112	1
14	ENST00000456328	sample12113	7
15	ENST00000456328	sample12114	7
16	ENST00000456328	sample12115	3
17	ENST00000456328	sample12117	24
18	ENST00000456328	sample12118	5
19	ENST00000456328	sample12119	9
20	ENST00000456328	sample12120	11
21	ENST00000456328	sample12121	11

Name: transcript_counts_table

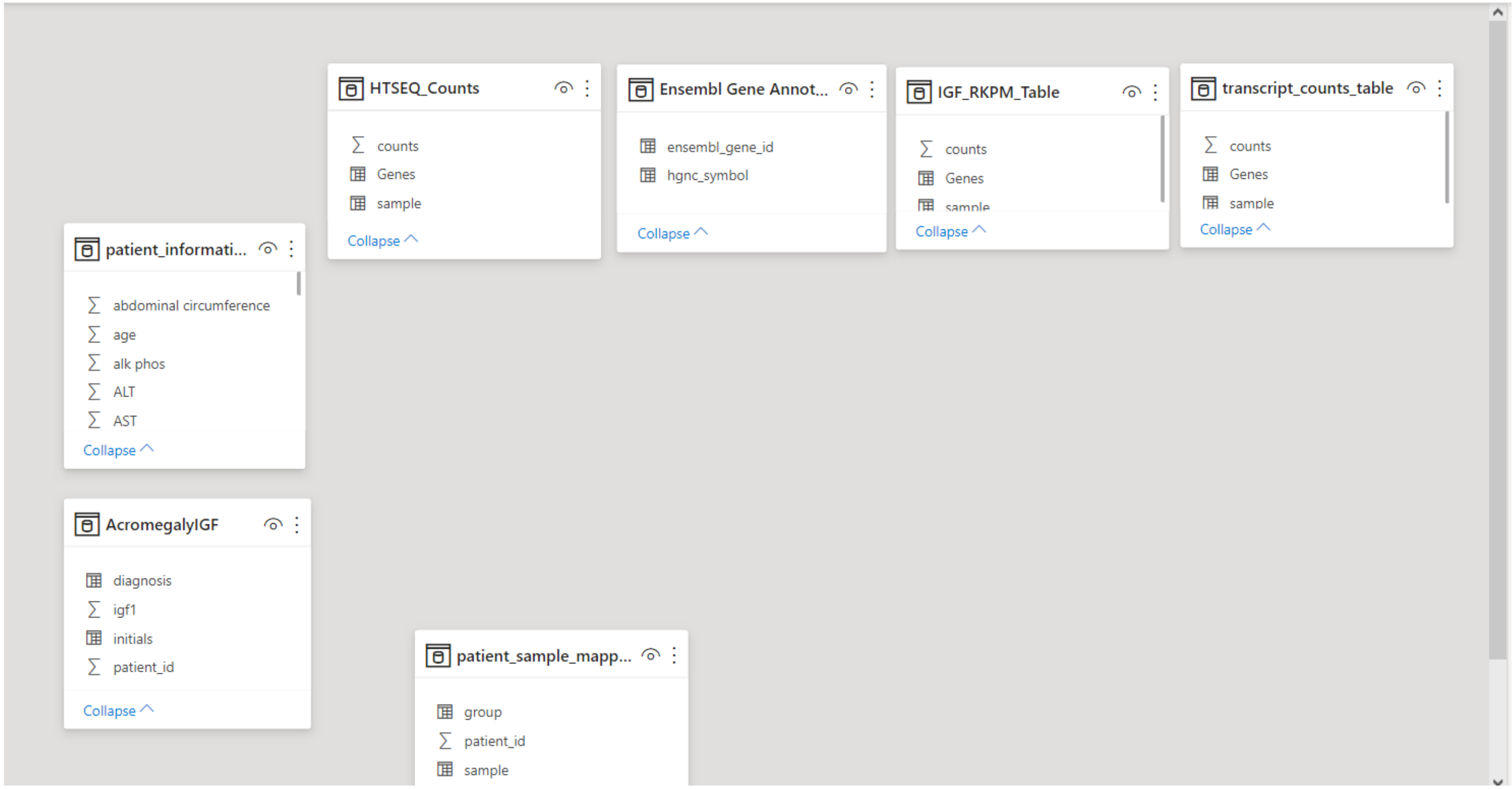
APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Renamed Columns
- UnPivotCols
- Renamecol1

3 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 16:37

The model view of the data table is as below.



HTSEQ_Counts
∑ counts
▦ Genes
▦ sample
Collapse ^

Ensembl Gene Annot...
▦ ensembl_gene_id
▦ hgnc_symbol
Collapse ^

IGF_RKPM_Table
∑ counts
▦ Genes
▦ sample
Collapse ^

transcript_counts_table
∑ counts
▦ Genes
▦ sample
Collapse ^

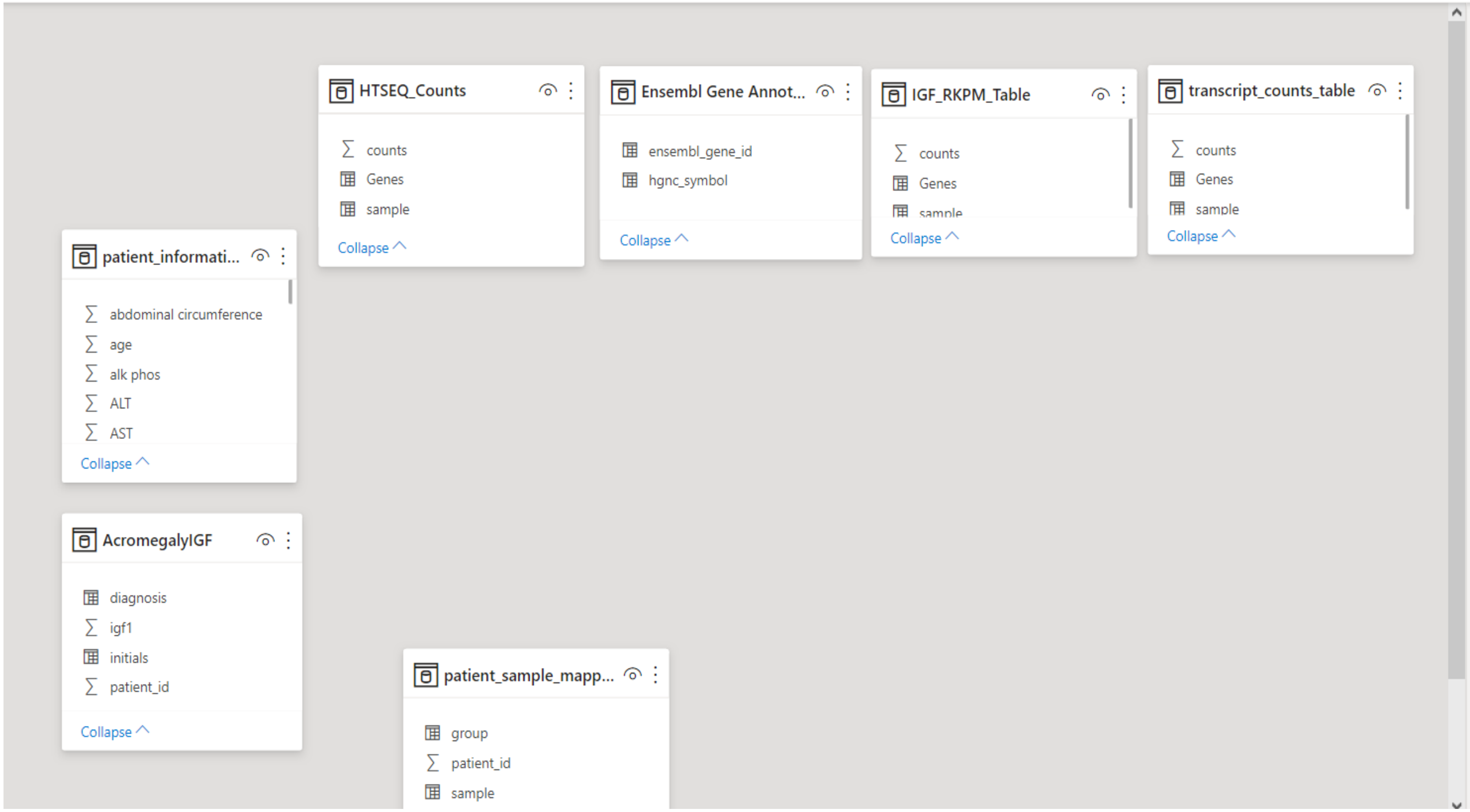
patient_informati...
∑ abdominal circumference
∑ age
∑ alk phos
∑ ALT
∑ AST
Collapse ^

AcromegalyIGF
▦ diagnosis
∑ igf1
▦ initials
∑ patient_id
Collapse ^

patient_sample_mapp...
▦ group
∑ patient_id
▦ sample

2. Star Schema / Snowflake Schema – Facts and Dimensions

On importing the data and processing the data in the tables, the Model view of the data is shown below.



HTSEQ_Counts
- ∑ counts
- ▦ Genes
- ▦ sample

Ensembl Gene Annot...
- ▦ ensembl_gene_id
- ▦ hgnc_symbol

IGF_RKPM_Table
- ∑ counts
- ▦ Genes
- ▦ sample

transcript_counts_table
- ∑ counts
- ▦ Genes
- ▦ sample

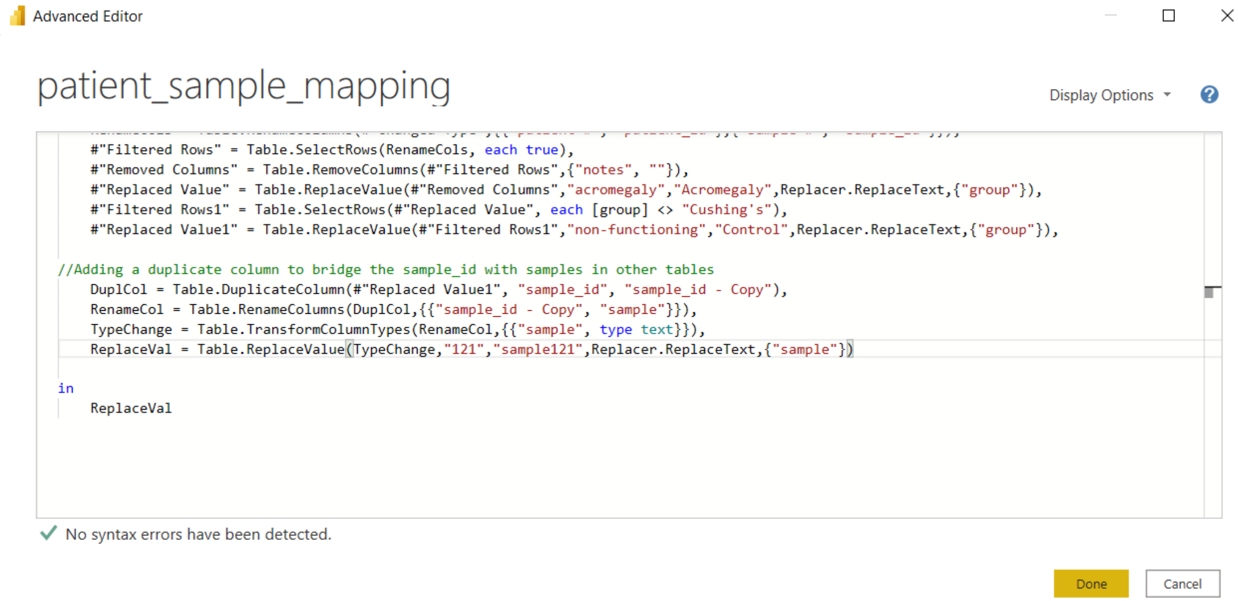
patient_informati...
- ∑ abdominal circumference
- ∑ age
- ∑ alk phos
- ∑ ALT
- ∑ AST

AcromegalyIGF
- ▦ diagnosis
- ∑ igf1
- ▦ initials
- ∑ patient_id

patient_sample_mapp...
- ▦ group
- ∑ patient_id
- ▦ sample

From the model view, it is evident that **the patient_sample_maping** table is the connecting table between all the other tables. Patient information and acromegaly patient-related data are stored in the tables with **patient_id** as the key. The gene_id, sequence counts in the other tables are related with sample_id as the key value.

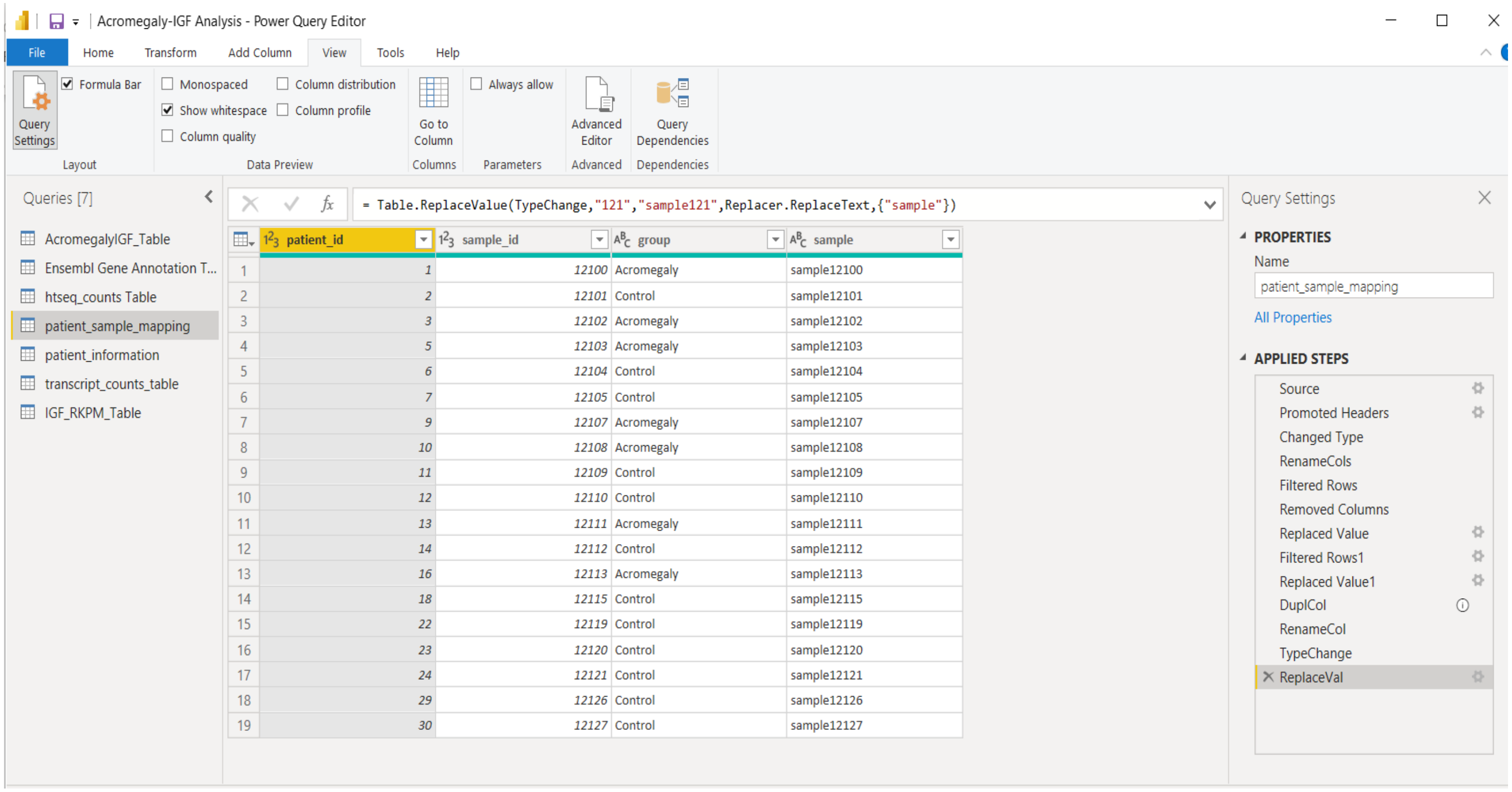
To connect the samples in **htseq_counts, IGF_RKPM_Counts,** and **transcript_count**s with the patients, we need to create a custom column. Create a custom column "**sample**" in the **patient_sample_mapping**" table using the formula below.



patient_sample_mapping

```
    #"Filtered Rows" = Table.SelectRows(RenameCols, each true),
    #"Removed Columns" = Table.RemoveColumns(#"Filtered Rows",{"notes", ""}),
    #"Replaced Value" = Table.ReplaceValue(#"Removed Columns","acromegaly","Acromegaly",Replacer.ReplaceText,{"group"}),
    #"Filtered Rows1" = Table.SelectRows(#"Replaced Value", each [group] <> "Cushing's"),
    #"Replaced Value1" = Table.ReplaceValue(#"Filtered Rows1","non-functioning","Control",Replacer.ReplaceText,{"group"}),

//Adding a duplicate column to bridge the sample_id with samples in other tables
    DuplCol = Table.DuplicateColumn(#"Replaced Value1", "sample_id", "sample_id - Copy"),
    RenameCol = Table.RenameColumns(DuplCol,{{"sample_id - Copy", "sample"}}),
    TypeChange = Table.TransformColumnTypes(RenameCol,{{"sample", type text}}),
    ReplaceVal = Table.ReplaceValue(TypeChange,"121","sample121",Replacer.ReplaceText,{"sample"})

in
    ReplaceVal
```

✓ No syntax errors have been detected.



☑ Formula Bar
□ Monospaced
☑ Show whitespace
□ Column quality
□ Column distribution
□ Column profile
□ Always allow

Queries [7]

- AcromegalyIGF_Table
- Ensembl Gene Annotation T...
- htseq_counts Table
- patient_sample_mapping
- patient_information
- transcript_counts_table
- IGF_RKPM_Table

```
= Table.ReplaceValue(TypeChange,"121","sample121",Replacer.ReplaceText,{"sample"})
```

	patient_id	sample_id	group	sample
1	1	12100	Acromegaly	sample12100
2	2	12101	Control	sample12101
3	3	12102	Acromegaly	sample12102
4	5	12103	Acromegaly	sample12103
5	6	12104	Control	sample12104
6	7	12105	Control	sample12105
7	9	12107	Acromegaly	sample12107
8	10	12108	Acromegaly	sample12108
9	11	12109	Control	sample12109
10	12	12110	Control	sample12110
11	13	12111	Acromegaly	sample12111
12	14	12112	Control	sample12112
13	16	12113	Acromegaly	sample12113
14	18	12115	Control	sample12115
15	22	12119	Control	sample12119
16	23	12120	Control	sample12120
17	24	12121	Control	sample12121
18	29	12126	Control	sample12126
19	30	12127	Control	sample12127

Query Settings

PROPERTIES

Name

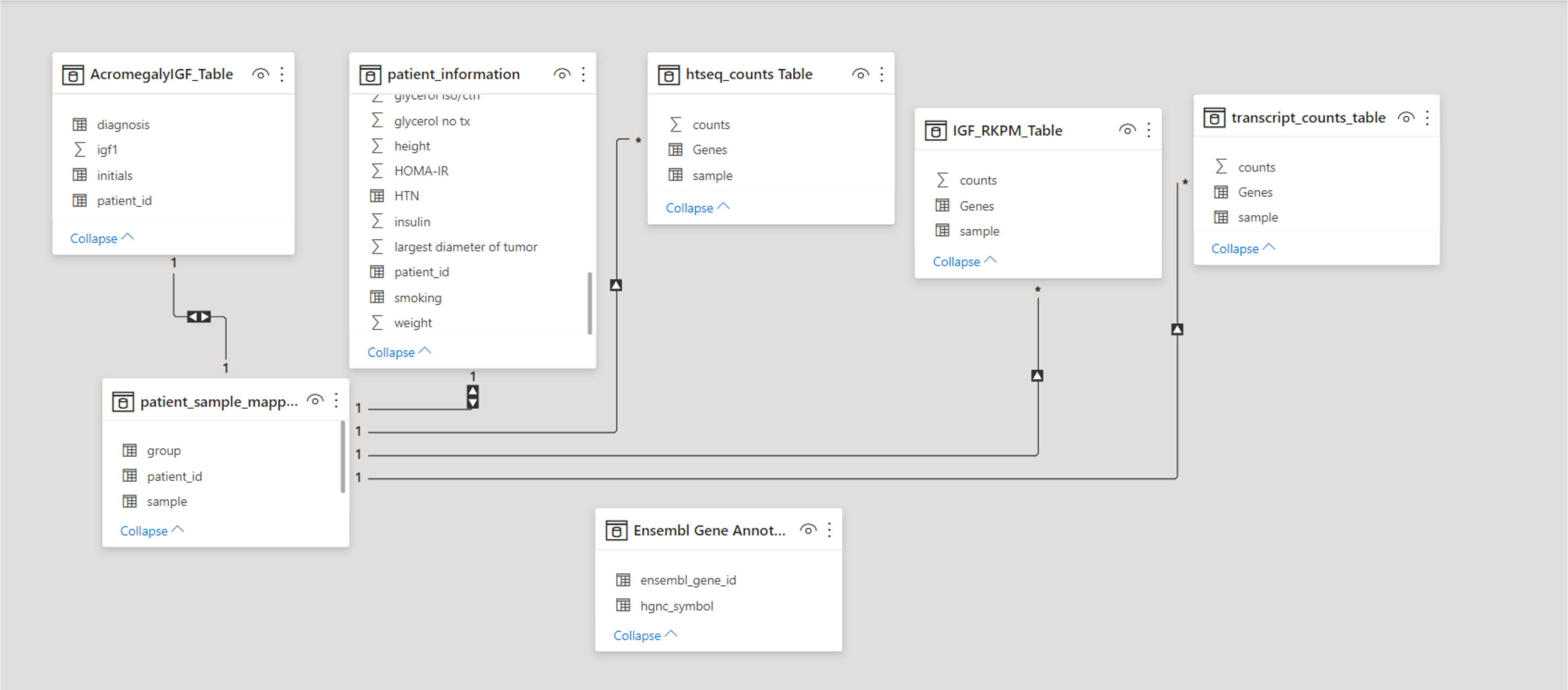
patient_sample_mapping

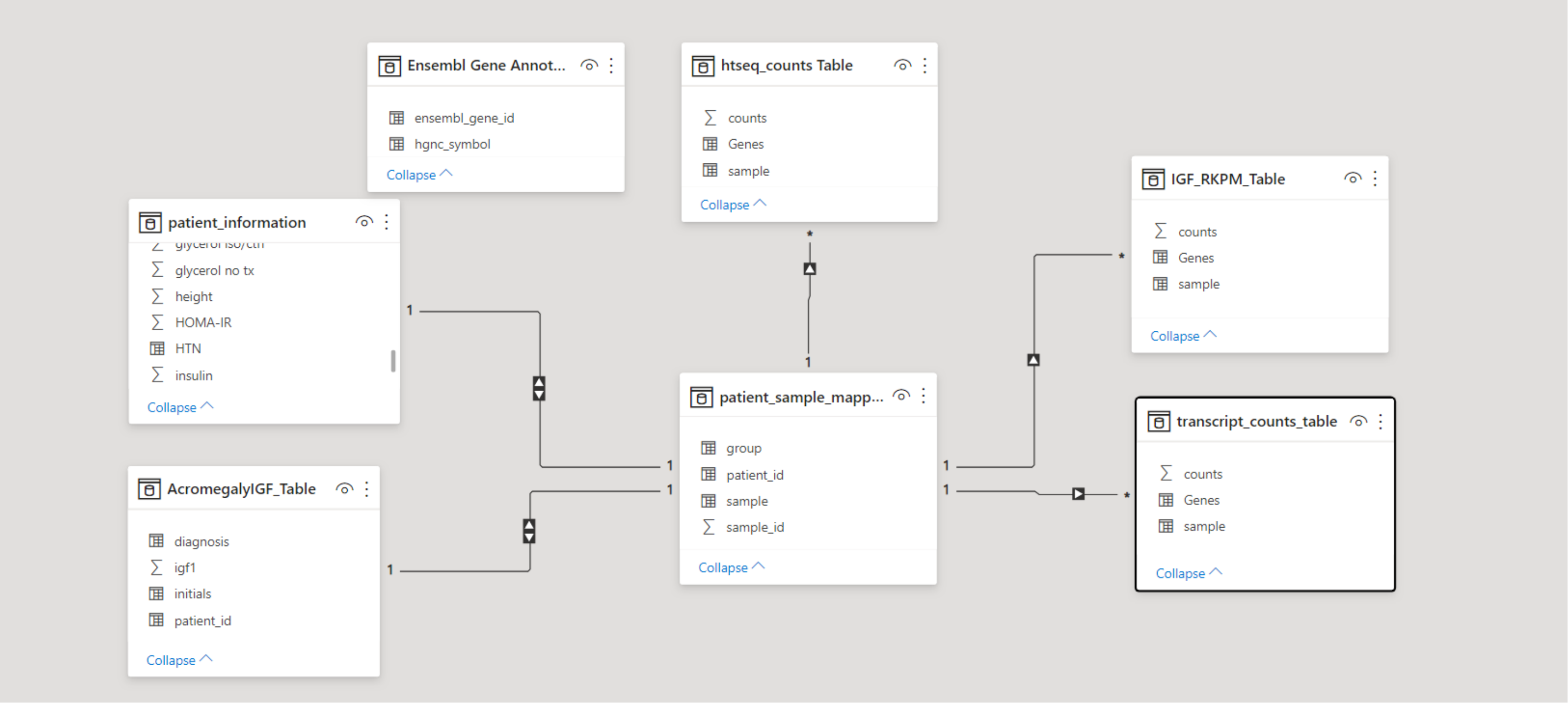
All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- RenameCols
- Filtered Rows
- Removed Columns
- Replaced Value
- Filtered Rows1
- Replaced Value1
- DuplCol
- RenameCol
- TypeChange
- ReplaceVal

On creating the new column, the tables can be related using **sample_id** and **patient_id** as below, while avoiding the many to many relationships.





The schema looks like a star schema (before connecting Ensembl Gene Annotation Table into the model).

2.1. Dimension Tables

Patient_information Table – stores all the patient-related data with **patient_**id as the key.

AcromegalyIGF_Table – Contains the information of acromegaly patients and their IGF levels with **patient_**id as key

Htseq_counts Table – maps the htsequence gene ids to sample counts with the **sample** as its key column

IGF_RKPM_Table – stores the IGF RKPM counts with the **sample** as the key column

Transcript_counts_table – Contains the information of gene id and counts **sample** as the key

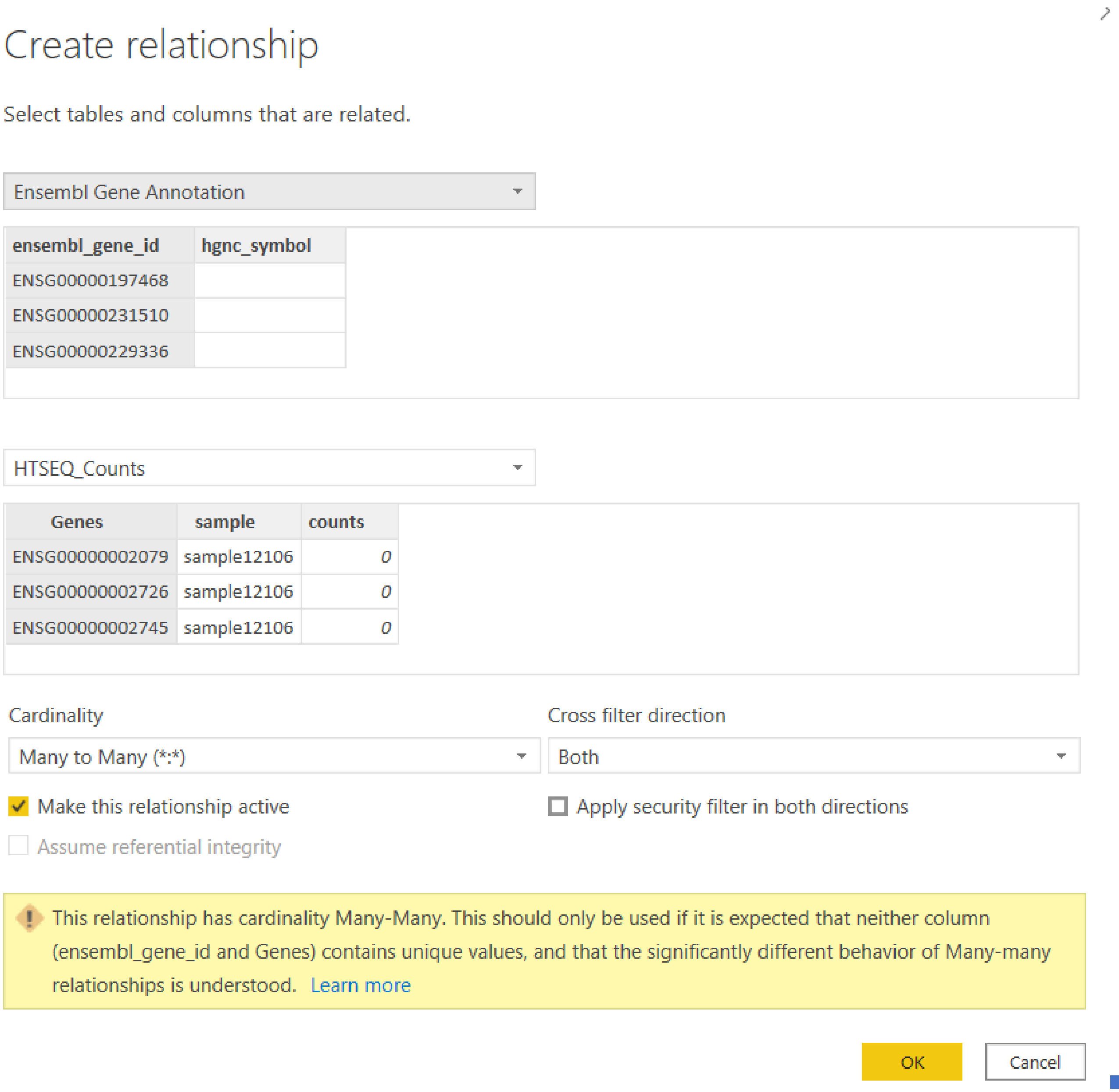
2.2. Fact Table

The fact table "patient_sample_mapping" doesn't store any information relating to any events, measures, or other information. This table only acts as a mapping bridge between other dimension tables. Patient_sample_mapping table has only columns that serve as a key in the dimension table. Hence, we can say this is a **fact-less fact table**.

Until we establish the connection of Ensembl Gene Annotation table, the data model replicates **Star Schema** with Dimension tables and Fact less fact table,

2.3. Avoid Many to Many Relationships

To connect the **Ensembl Gene Annotation table** we have only one related column is "Genes" in **htseq_counts Table**. And we only need IGF-related Gene Information. When we connect these tables, it shows many to many relationships as shown in the screenshot below.



Create relationship

Select tables and columns that are related.

Ensembl Gene Annotation

ensembl_gene_id	hgnc_symbol
ENSG00000197468	
ENSG00000231510	
ENSG00000229336	

HTSEQ_Counts

Genes	sample	counts
ENSG00000002079	sample12106	0
ENSG00000002726	sample12106	0
ENSG00000002745	sample12106	0

Cardinality: Many to Many (*:*)

Cross filter direction: Both

☑ Make this relationship active

□ Apply security filter in both directions

□ Assume referential integrity

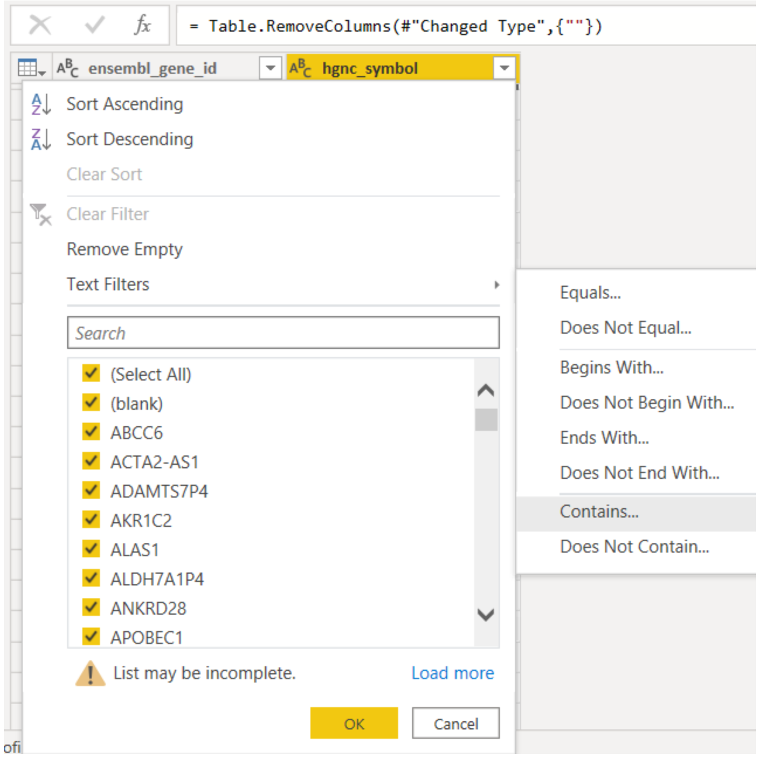
This relationship has cardinality Many-Many. This should only be used if it is expected that neither column (ensembl_gene_id and Genes) contains unique values, and that the significantly different behavior of Many-many relationships is understood. Learn more

To avoid this many to many and connect the tables we need to follow 3 steps as below.

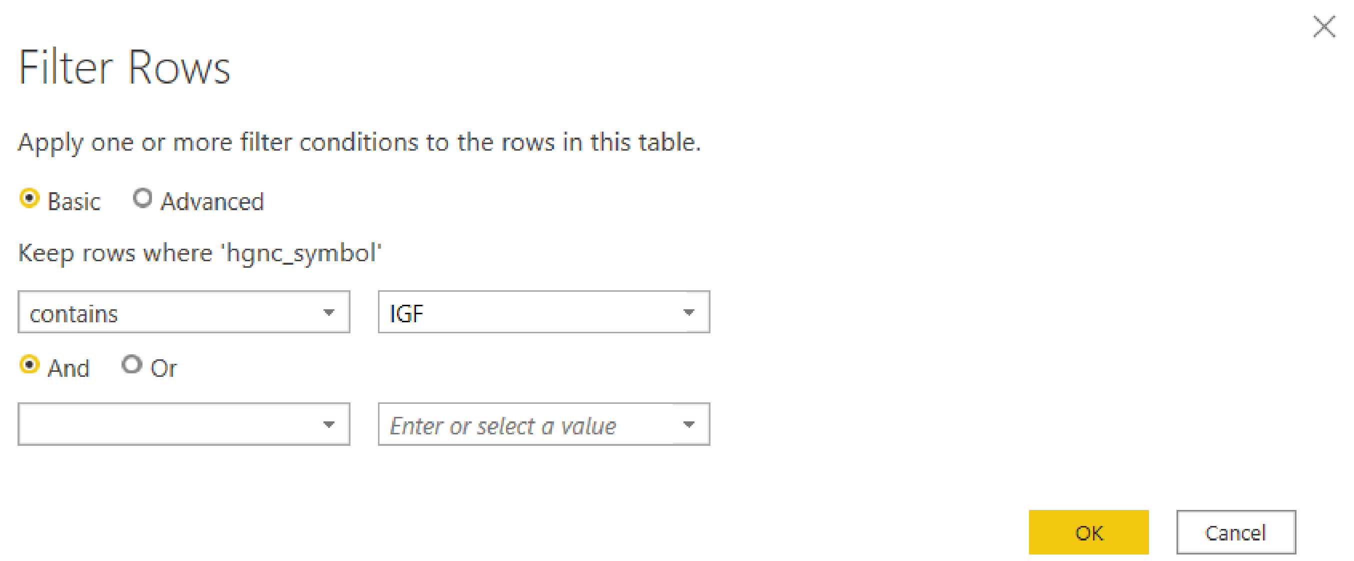
1. Filter Rows: We need only IGF gene-related Data, so we filter the Ensembl Gene Annotation table with hgnc_symbol containing “igf”
2. Merge Queries: merge htseq_counts Table with patient_sample_mapping and create a new column with diagnosis “group” details.
3. Create a New Relationship :

2.4. Filter Rows

We need only IGF gene-related Data, so we filter the Ensembl Gene Annotation table with hgnc_symbol containing “igf”



```
= Table.RemoveColumns(#"Changed Type",{""})
```



Filter Rows

Apply one or more filter conditions to the rows in this table.

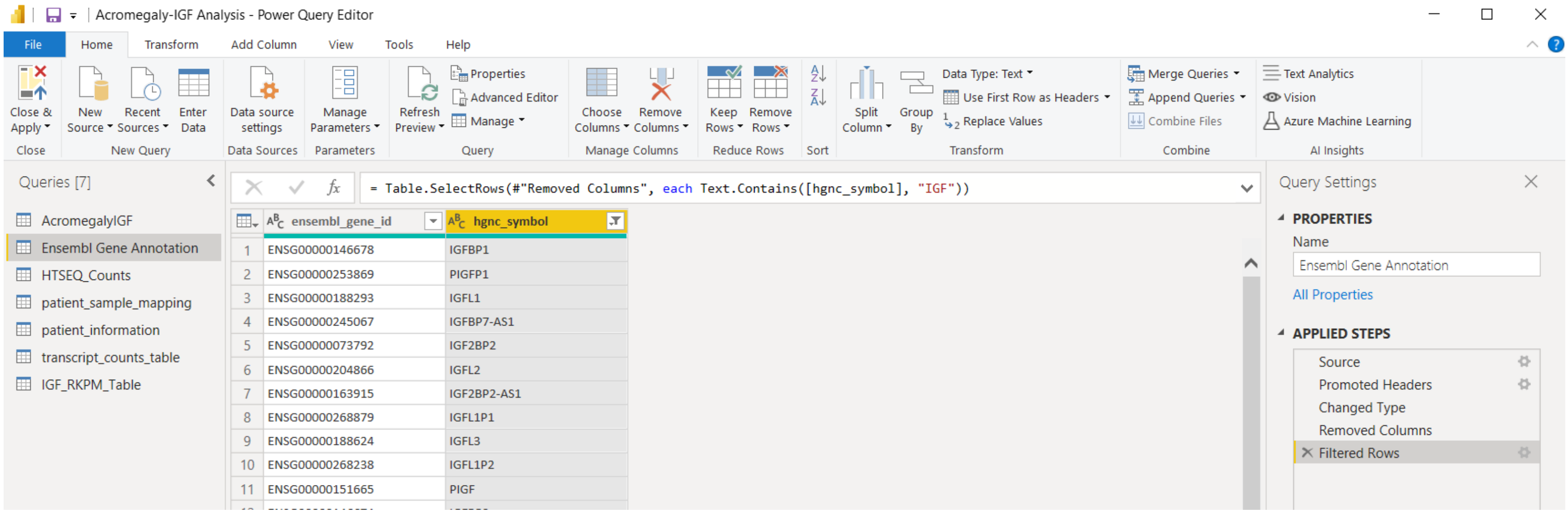
● Basic ○ Advanced

Keep rows where 'hgnc_symbol'

contains IGF

● And ○ Or

Enter or select a value



Acromegaly-IGF Analysis - Power Query Editor

```
= Table.SelectRows(#"Removed Columns", each Text.Contains([hgnc_symbol], "IGF"))
```

	ensembl_gene_id	hgnc_symbol
1	ENSG00000146678	IGFBP1
2	ENSG00000253869	PIGFP1
3	ENSG00000188293	IGFL1
4	ENSG00000245067	IGFBP7-AS1
5	ENSG00000073792	IGF2BP2
6	ENSG00000204866	IGFL2
7	ENSG00000163915	IGF2BP2-AS1
8	ENSG00000268879	IGFL1P1
9	ENSG00000188624	IGFL3
10	ENSG00000268238	IGFL1P2
11	ENSG00000151665	PIGF

Query Settings

PROPERTIES

Name

Ensembl Gene Annotation

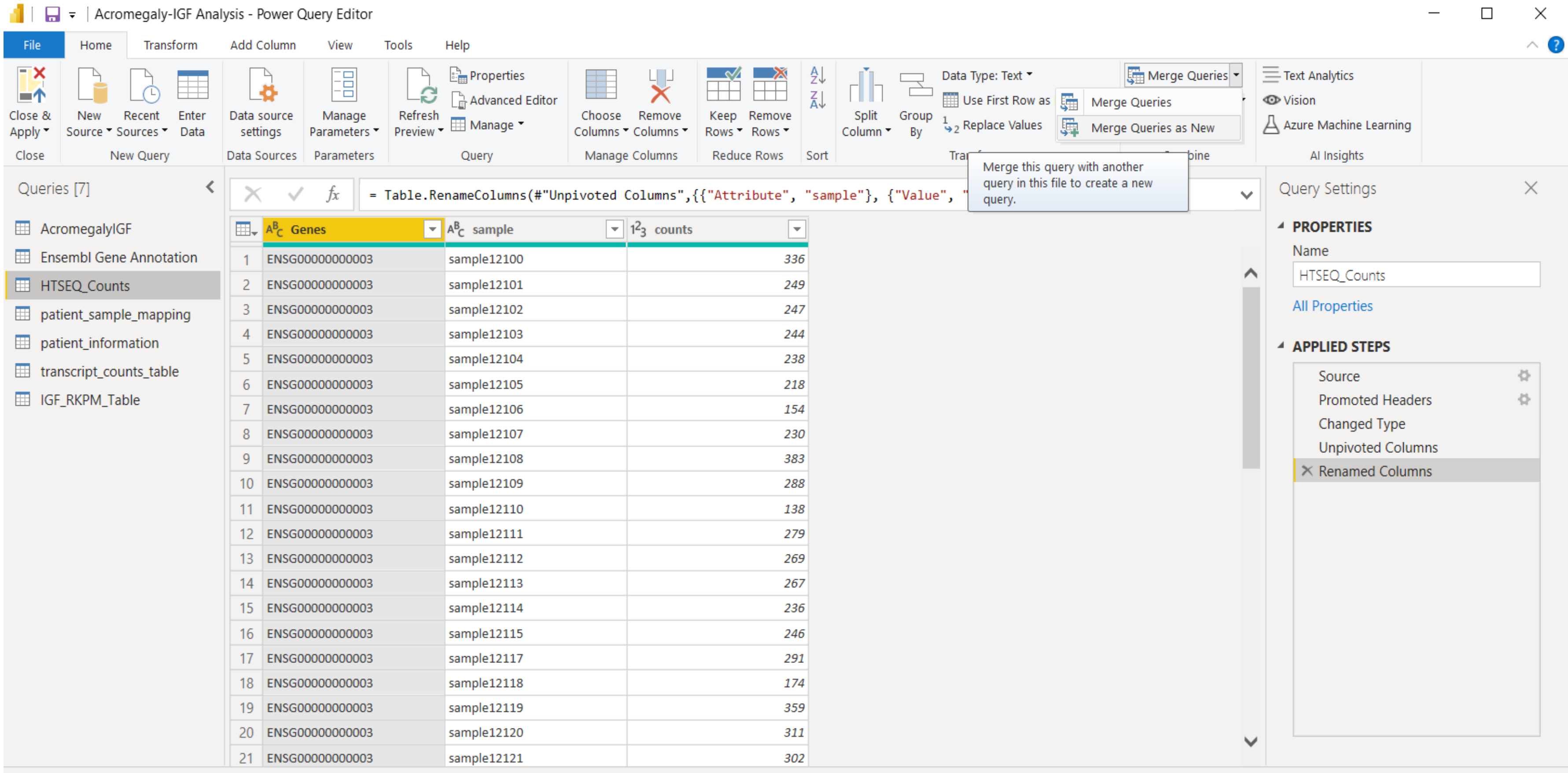
All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Removed Columns
- Filtered Rows

2.5. Merge Queries

Merging htseq_counts Table with patient_sample_mapping and create a new column with diagnosis “group” details.



Merge this query with another query in this file to create a new query.

Queries [7]

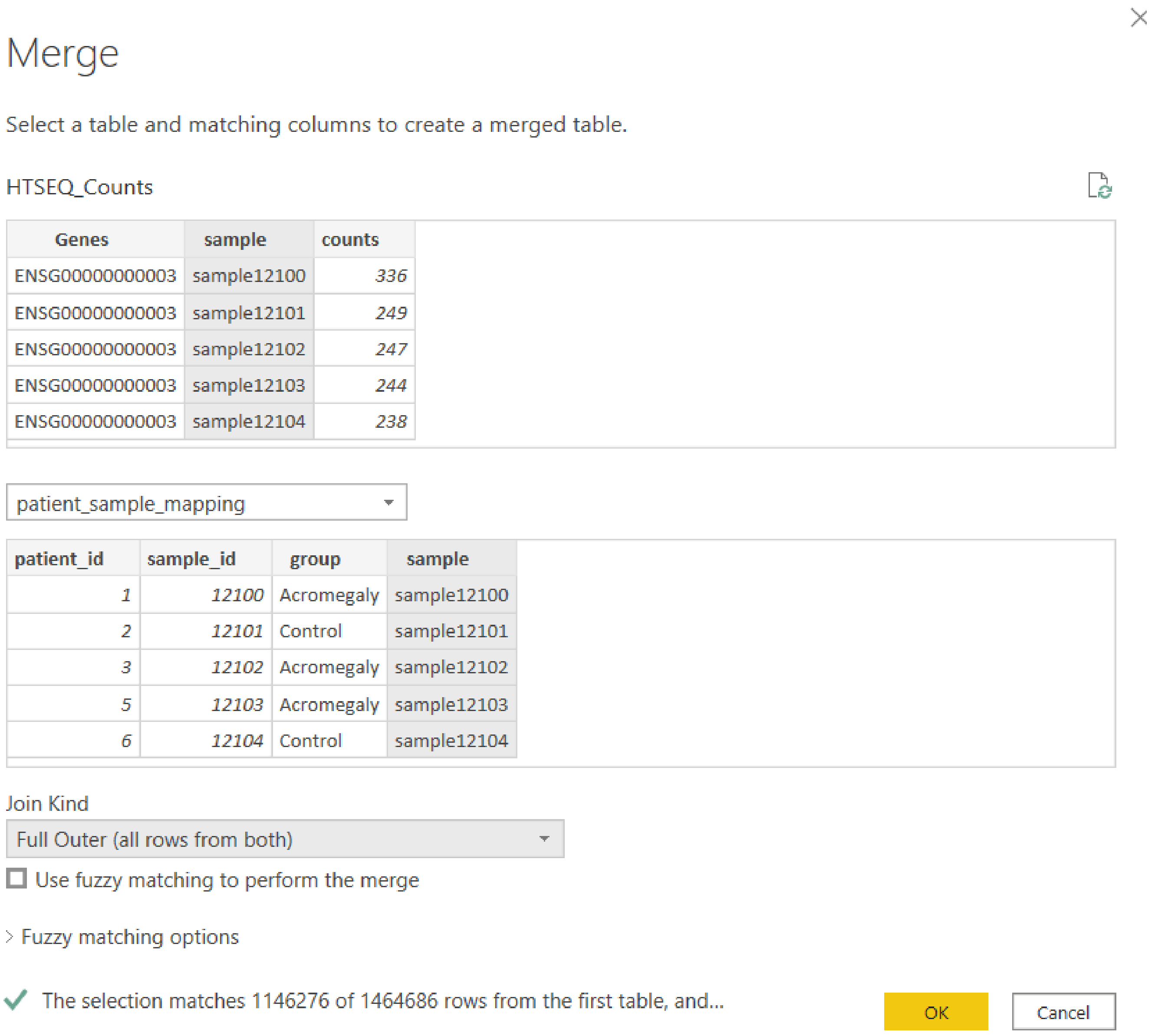
- AcromegalyIGF
- Ensembl Gene Annotation
- HTSEQ_Counts
- patient_sample_mapping
- patient_information
- transcript_counts_table
- IGF_RKPM_Table

```
= Table.RenameColumns(#"Unpivoted Columns",{{"Attribute", "sample"}, {"Value", "
```

	Genes	sample	counts
1	ENSG00000000003	sample12100	336
2	ENSG00000000003	sample12101	249
3	ENSG00000000003	sample12102	247
4	ENSG00000000003	sample12103	244
5	ENSG00000000003	sample12104	238
6	ENSG00000000003	sample12105	218
7	ENSG00000000003	sample12106	154
8	ENSG00000000003	sample12107	230
9	ENSG00000000003	sample12108	383
10	ENSG00000000003	sample12109	288
11	ENSG00000000003	sample12110	138
12	ENSG00000000003	sample12111	279
13	ENSG00000000003	sample12112	269
14	ENSG00000000003	sample12113	267
15	ENSG00000000003	sample12114	236
16	ENSG00000000003	sample12115	246
17	ENSG00000000003	sample12117	291
18	ENSG00000000003	sample12118	174
19	ENSG00000000003	sample12119	359
20	ENSG00000000003	sample12120	311
21	ENSG00000000003	sample12121	302

Name: HTSEQ_Counts

- Source
- Promoted Headers
- Changed Type
- Unpivoted Columns
- Renamed Columns



Merge

Select a table and matching columns to create a merged table.

HTSEQ_Counts

Genes	sample	counts
ENSG00000000003	sample12100	336
ENSG00000000003	sample12101	249
ENSG00000000003	sample12102	247
ENSG00000000003	sample12103	244
ENSG00000000003	sample12104	238

patient_sample_mapping

patient_id	sample_id	group	sample
1	12100	Acromegaly	sample12100
2	12101	Control	sample12101
3	12102	Acromegaly	sample12102
5	12103	Acromegaly	sample12103
6	12104	Control	sample12104

Join Kind

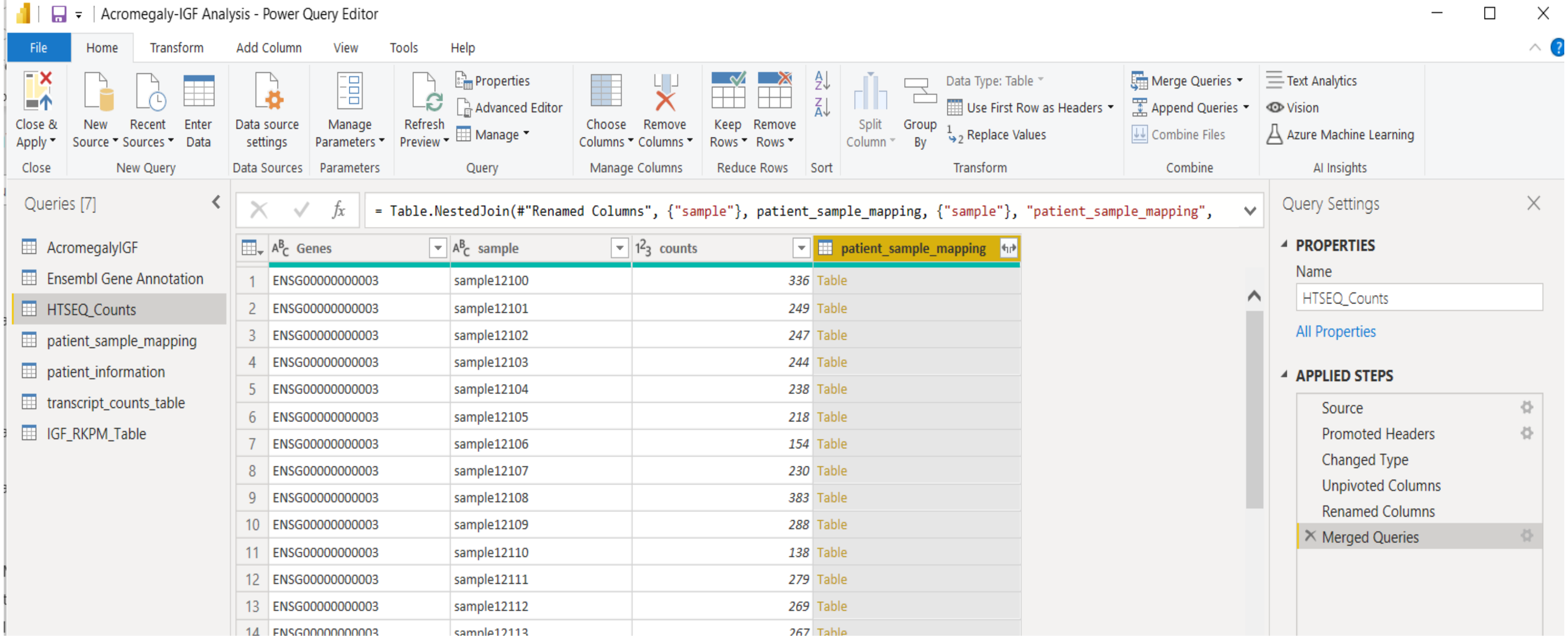
Full Outer (all rows from both)

□ Use fuzzy matching to perform the merge

✓ The selection matches 1146276 of 1464686 rows from the first table, and...

Selecting group Column from

Selecting **“group”** column from the resulting table after merge.



Queries [7]

- AcromegalyIGF
- Ensembl Gene Annotation
- HTSEQ_Counts
- patient_sample_mapping
- patient_information
- transcript_counts_table
- IGF_RKPM_Table

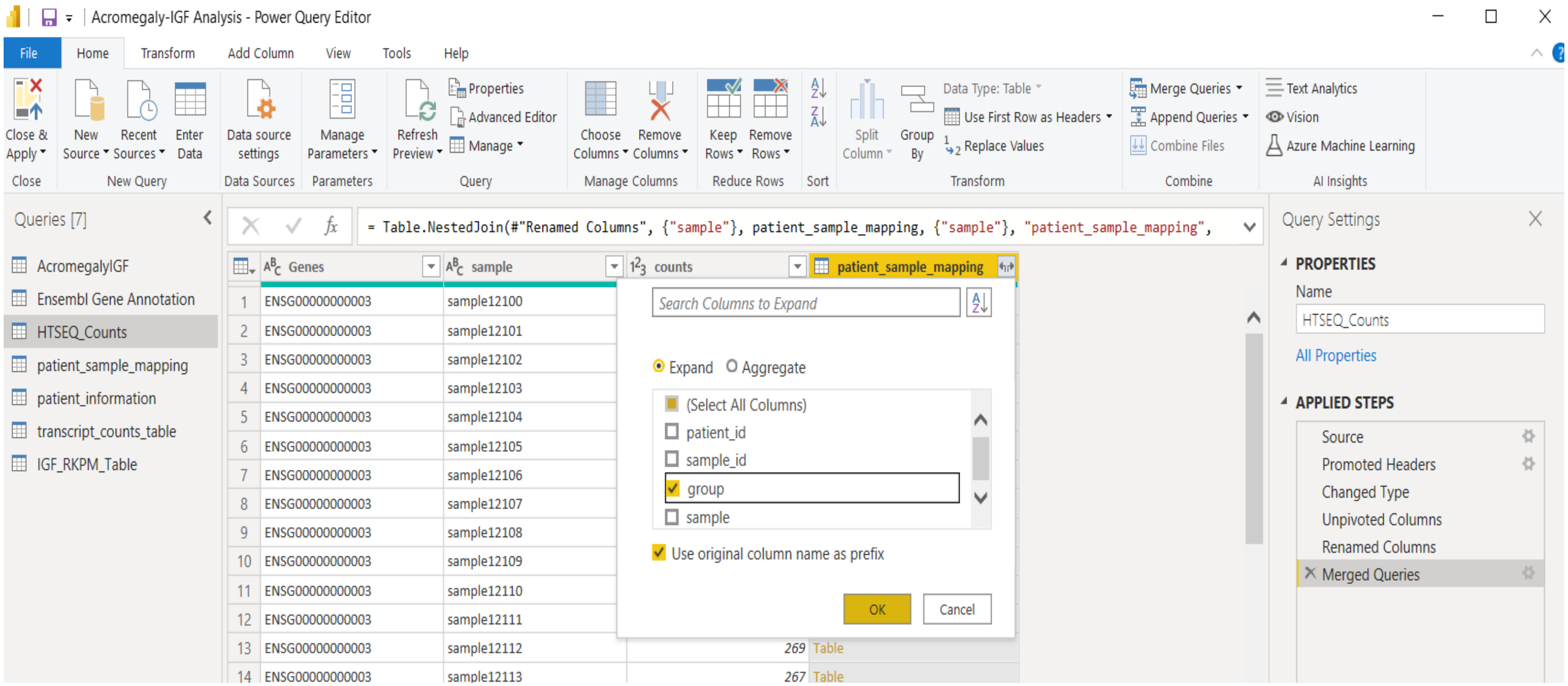
```
= Table.NestedJoin(#"Renamed Columns", {"sample"}, patient_sample_mapping, {"sample"}, "patient_sample_mapping",
```

	Genes	sample	counts	patient_sample_mapping
1	ENSG00000000003	sample12100	336	Table
2	ENSG00000000003	sample12101	249	Table
3	ENSG00000000003	sample12102	247	Table
4	ENSG00000000003	sample12103	244	Table
5	ENSG00000000003	sample12104	238	Table
6	ENSG00000000003	sample12105	218	Table
7	ENSG00000000003	sample12106	154	Table
8	ENSG00000000003	sample12107	230	Table
9	ENSG00000000003	sample12108	383	Table
10	ENSG00000000003	sample12109	288	Table
11	ENSG00000000003	sample12110	138	Table
12	ENSG00000000003	sample12111	279	Table
13	ENSG00000000003	sample12112	269	Table
14	ENSG00000000003	sample12113	267	Table

Name: HTSEQ_Counts

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Unpivoted Columns
- Renamed Columns
- Merged Queries



Queries [7]

- AcromegalyIGF
- Ensembl Gene Annotation
- HTSEQ_Counts
- patient_sample_mapping
- patient_information
- transcript_counts_table
- IGF_RKPM_Table

```
= Table.NestedJoin(#"Renamed Columns", {"sample"}, patient_sample_mapping, {"sample"}, "patient_sample_mapping",
```

	Genes	sample	counts	patient_sample_mapping
1	ENSG00000000003	sample12100		
2	ENSG00000000003	sample12101		
3	ENSG00000000003	sample12102		
4	ENSG00000000003	sample12103		
5	ENSG00000000003	sample12104		
6	ENSG00000000003	sample12105		
7	ENSG00000000003	sample12106		
8	ENSG00000000003	sample12107		
9	ENSG00000000003	sample12108		
10	ENSG00000000003	sample12109		
11	ENSG00000000003	sample12110		
12	ENSG00000000003	sample12111		
13	ENSG00000000003	sample12112	269	Table
14	ENSG00000000003	sample12113	267	Table

Search Columns to Expand

● Expand ○ Aggregate

- ▪ (Select All Columns)
- □ patient_id
- □ sample_id
- ☑ group
- □ sample

☑ Use original column name as prefix

OK Cancel

Query Settings

PROPERTIES

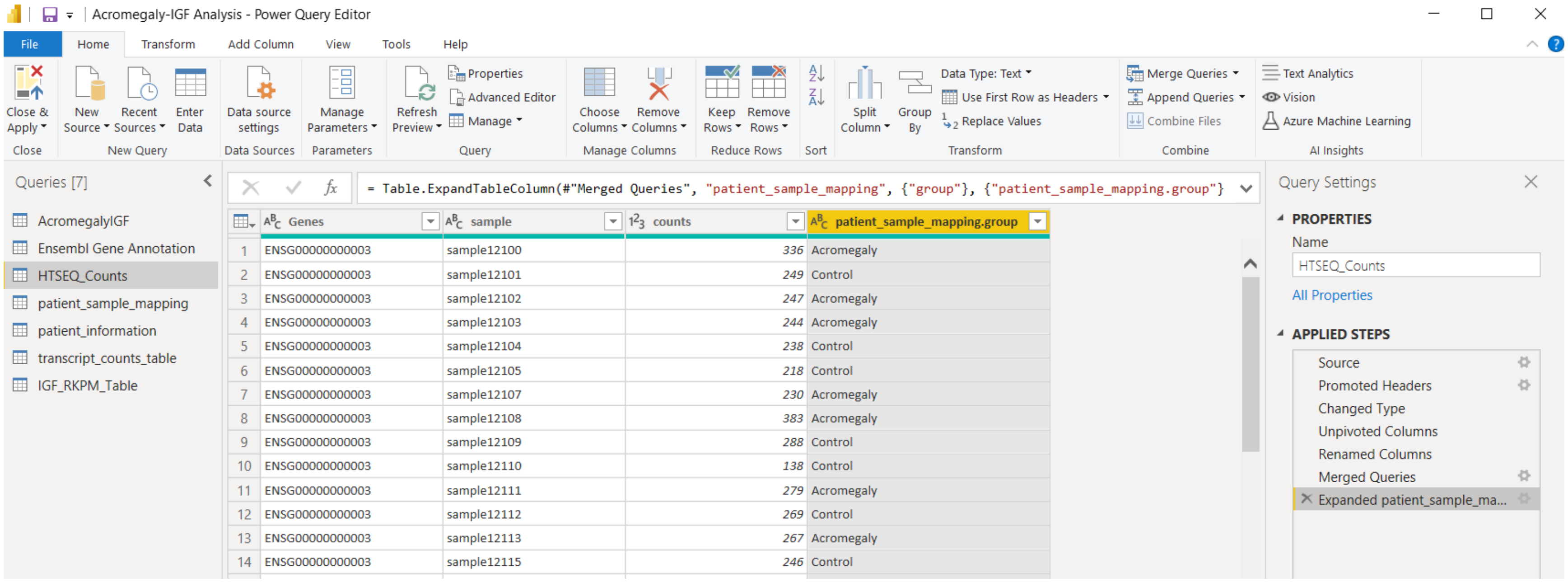
Name

HTSEQ_Counts

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Unpivoted Columns
- Renamed Columns
- Merged Queries



Queries [7]

- AcromegalyIGF
- Ensembl Gene Annotation
- HTSEQ_Counts
- patient_sample_mapping
- patient_information
- transcript_counts_table
- IGF_RKPM_Table

```
= Table.ExpandTableColumn(#"Merged Queries", "patient_sample_mapping", {"group"}, {"patient_sample_mapping.group"})
```

	Genes	sample	counts	patient_sample_mapping.group
1	ENSG00000000003	sample12100	336	Acromegaly
2	ENSG00000000003	sample12101	249	Control
3	ENSG00000000003	sample12102	247	Acromegaly
4	ENSG00000000003	sample12103	244	Acromegaly
5	ENSG00000000003	sample12104	238	Control
6	ENSG00000000003	sample12105	218	Control
7	ENSG00000000003	sample12107	230	Acromegaly
8	ENSG00000000003	sample12108	383	Acromegaly
9	ENSG00000000003	sample12109	288	Control
10	ENSG00000000003	sample12110	138	Control
11	ENSG00000000003	sample12111	279	Acromegaly
12	ENSG00000000003	sample12112	269	Control
13	ENSG00000000003	sample12113	267	Acromegaly
14	ENSG00000000003	sample12115	246	Control

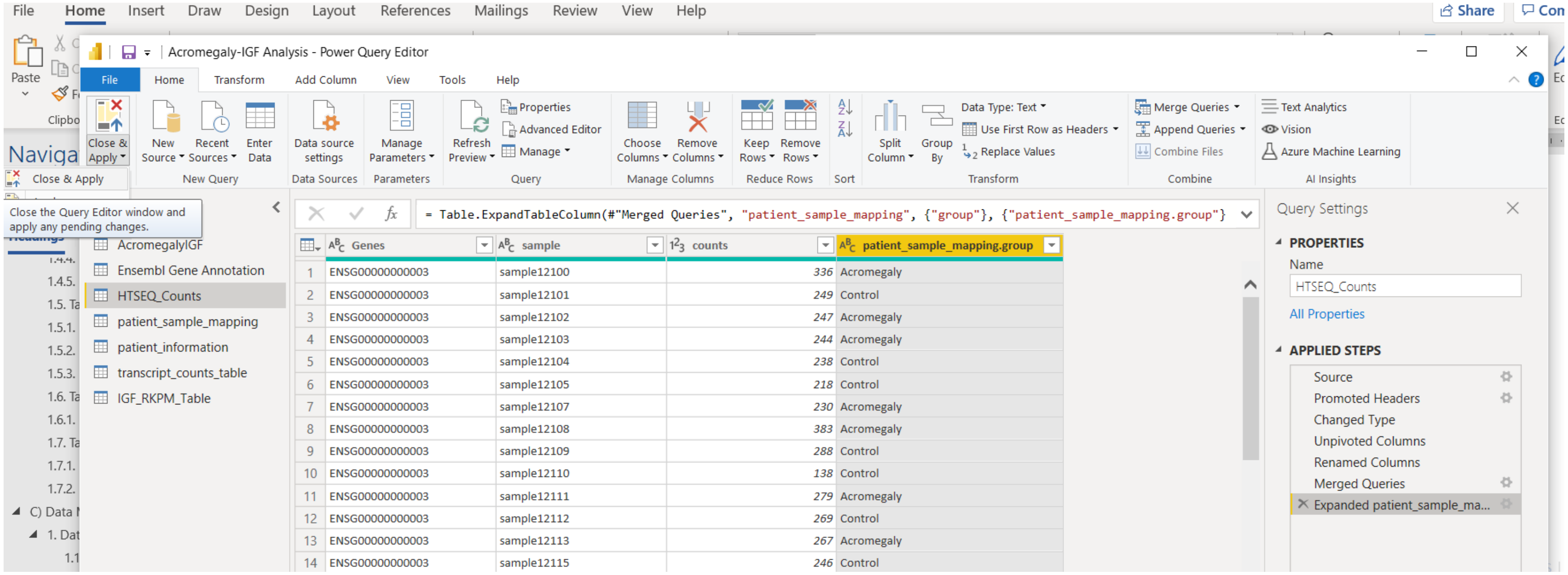
PROPERTIES

Name: HTSEQ_Counts

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type
- Unpivoted Columns
- Renamed Columns
- Merged Queries
- Expanded patient_sample_ma...



Acromegaly-IGF Analysis - Power Query Editor

Close & Apply

Close the Query Editor window and apply any pending changes.

- AcromegalyIGF
- Ensembl Gene Annotation
- HTSEQ_Counts
- patient_sample_mapping
- patient_information
- transcript_counts_table
- IGF_RKPM_Table

```
= Table.ExpandTableColumn(#"Merged Queries", "patient_sample_mapping", {"group"}, {"patient_sample_mapping.group"}
```

	Genes	sample	counts	patient_sample_mapping.group
1	ENSG00000000003	sample12100	336	Acromegaly
2	ENSG00000000003	sample12101	249	Control
3	ENSG00000000003	sample12102	247	Acromegaly
4	ENSG00000000003	sample12103	244	Acromegaly
5	ENSG00000000003	sample12104	238	Control
6	ENSG00000000003	sample12105	218	Control
7	ENSG00000000003	sample12107	230	Acromegaly
8	ENSG00000000003	sample12108	383	Acromegaly
9	ENSG00000000003	sample12109	288	Control
10	ENSG00000000003	sample12110	138	Control
11	ENSG00000000003	sample12111	279	Acromegaly
12	ENSG00000000003	sample12112	269	Control
13	ENSG00000000003	sample12113	267	Acromegaly
14	ENSG00000000003	sample12115	246	Control

Query Settings

PROPERTIES

Name

HTSEQ_Counts

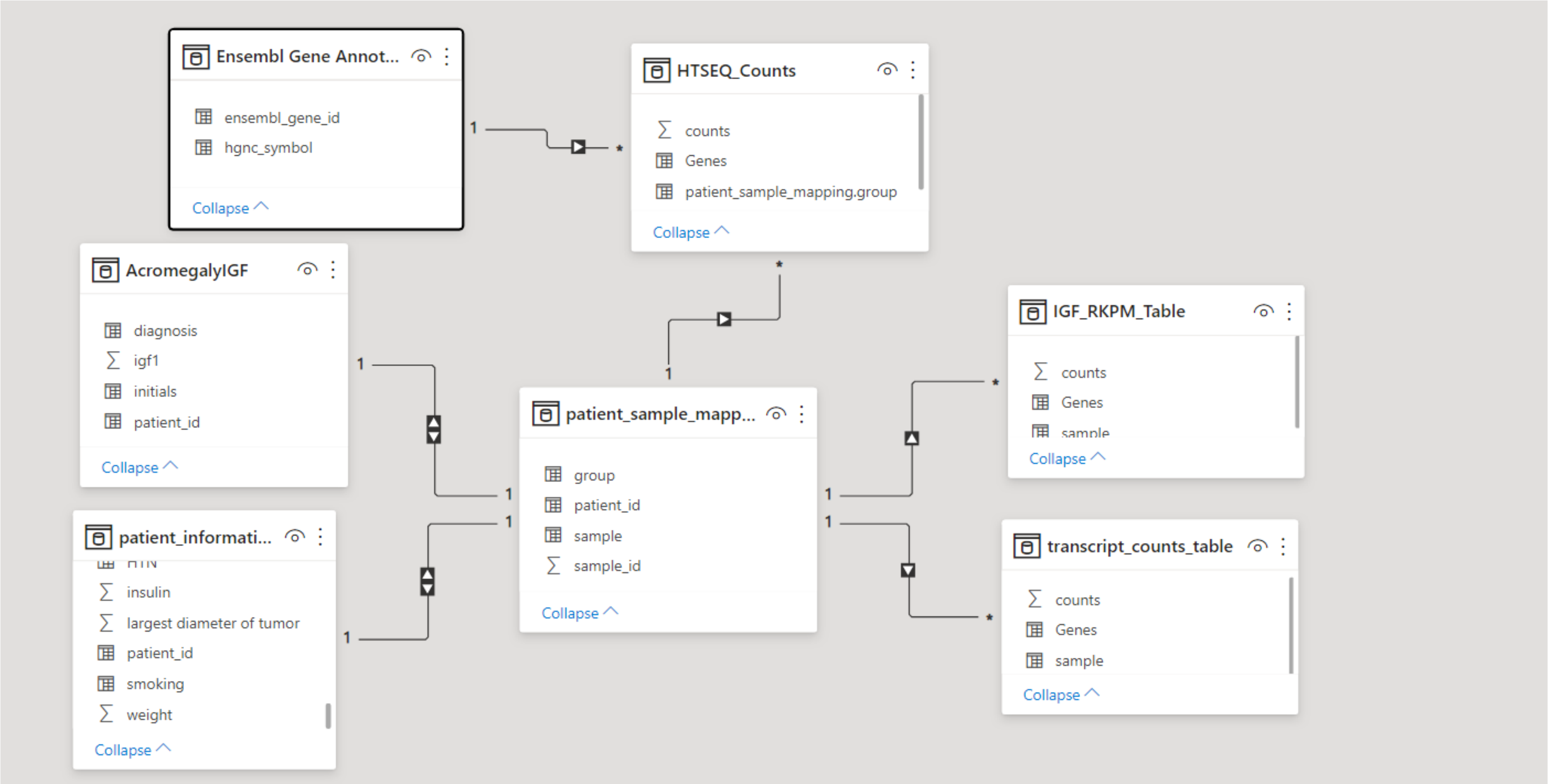
All Properties

APPLIED STEPS

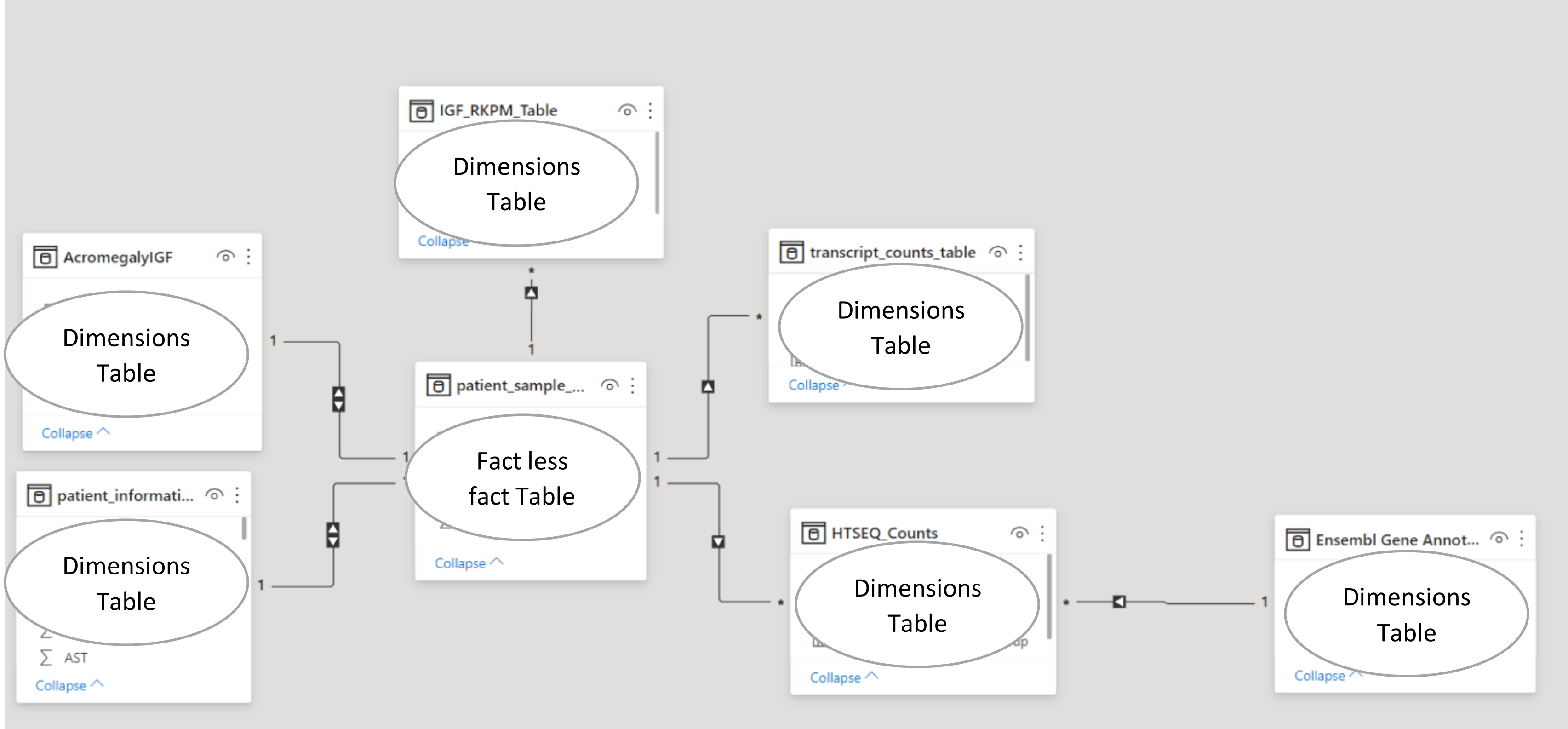
- Source
- Promoted Headers
- Changed Type
- Unpivoted Columns
- Renamed Columns
- Merged Queries
- Expanded patient_sample_ma...

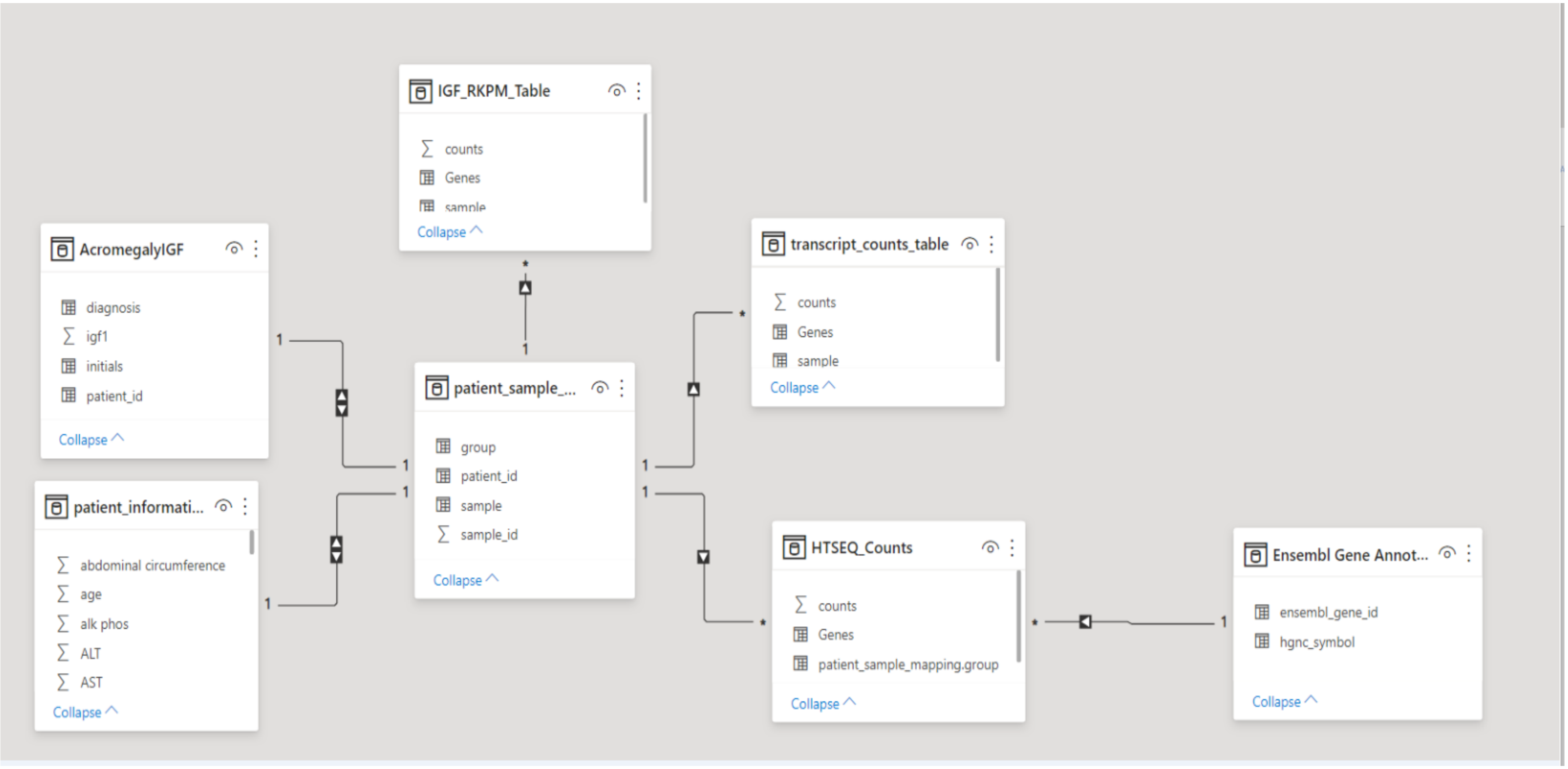
2.6. Create a New Relationship

Now we relate ensembl_gene_id to the Genes Column in the HTSEQ_Counts Table. This doesn't show any * to * relationships. Instead, it is mapped with one to many relations



Finally looking at the data model, we can confirm that it is a **snowflake schema**.


IGF_RKPM_Table
Dimensions Table
Collapse
AcromegalyIGF
Dimensions Table
Collapse
transcript_counts_table
Dimensions Table
Collapse
patient_sample_...
Fact less fact Table
Collapse
patient_informati...
Dimensions Table
Σ AST
Collapse
HTSEQ_Counts
Dimensions Table
Collapse
Ensembl Gene Annot...
Dimensions Table
Collapse
1
*
1
1
1
1
1
*
1
*
*
1


IGF_RKPM_Table
counts
Genes
sample
Collapse
AcromegalyIGF
diagnosis
igf1
initials
patient_id
Collapse
transcript_counts_table
counts
Genes
sample
Collapse
patient_sample_...
group
patient_id
sample
sample_id
Collapse
patient_informati...
abdominal circumference
age
alk phos
ALT
AST
Collapse
HTSEQ_Counts
counts
Genes
patient_sample_mapping.group
Collapse
Ensembl Gene Annot...
ensembl_gene_id
hgnc_symbol
Collapse
1
*
1
1
1
1
1
*
1
*
*
1