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

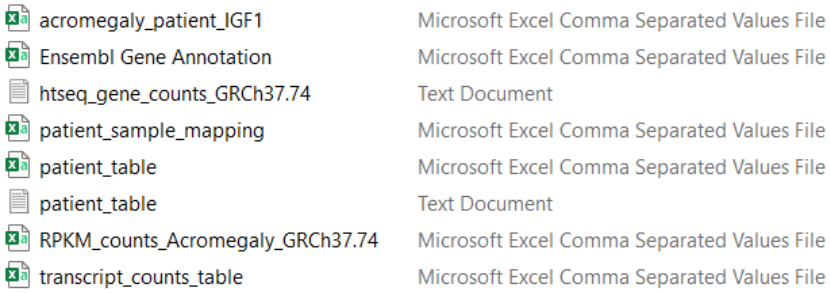


Figure 1 Raw Data Files

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.

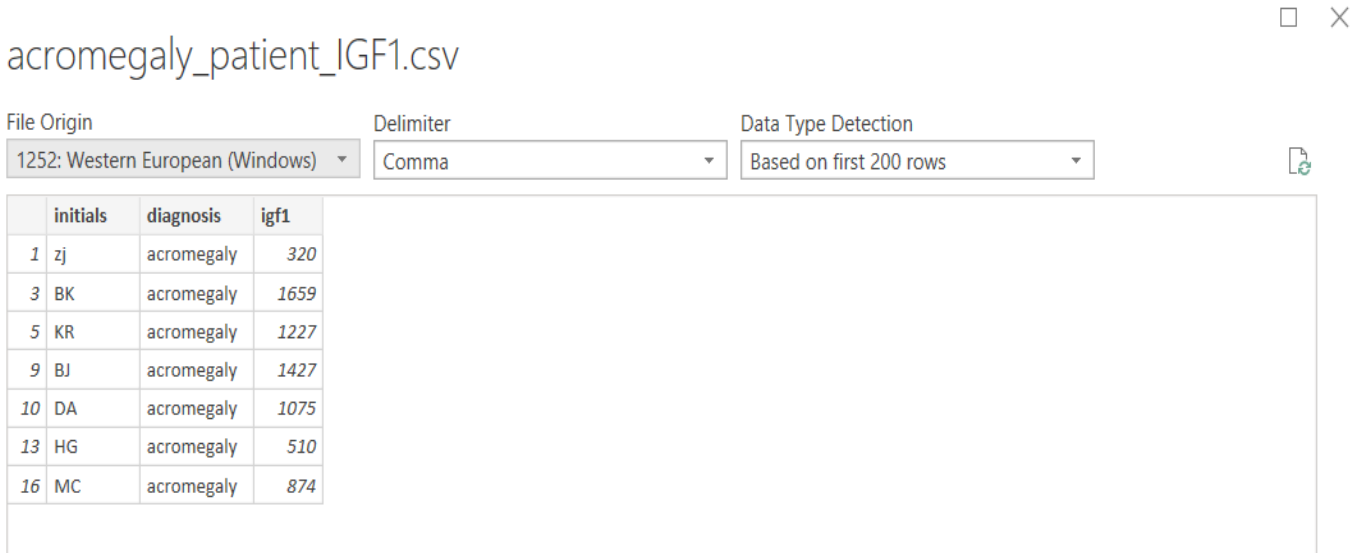
Table 2 Dataset Source

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.



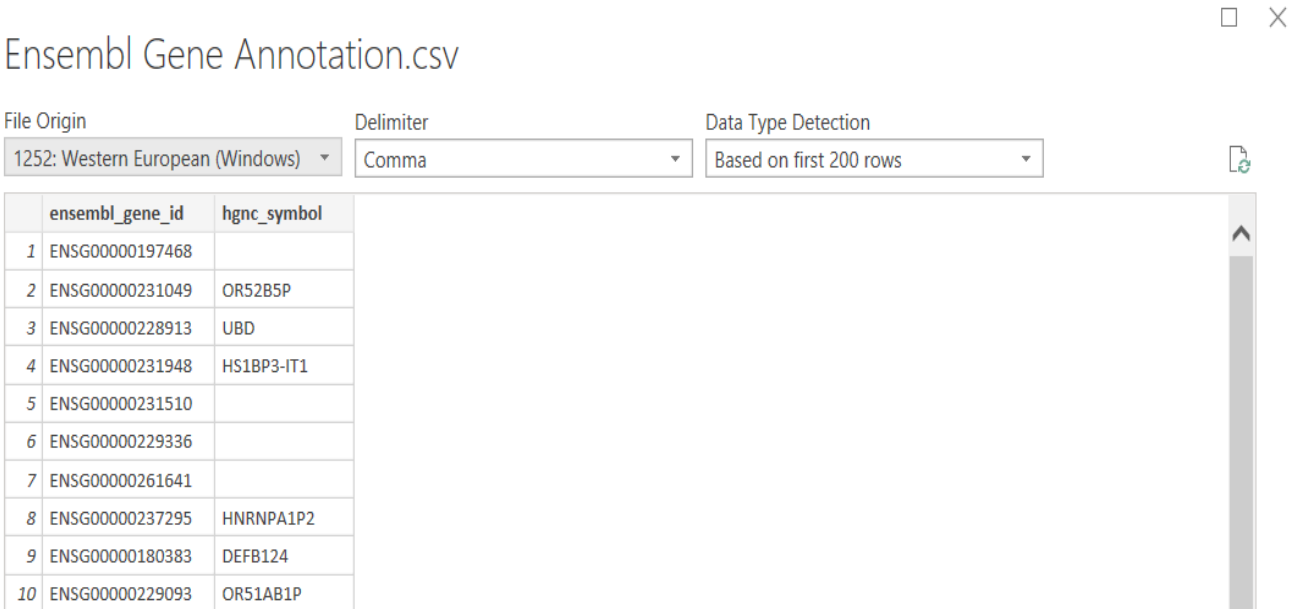
	initials	diagnosis	igf1
1	zj	acromegaly	320
3	BK	acromegaly	1659
5	KR	acromegaly	1227
9	BJ	acromegaly	1427
10	DA	acromegaly	1075
13	HG	acromegaly	510
16	MC	acromegaly	874

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_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

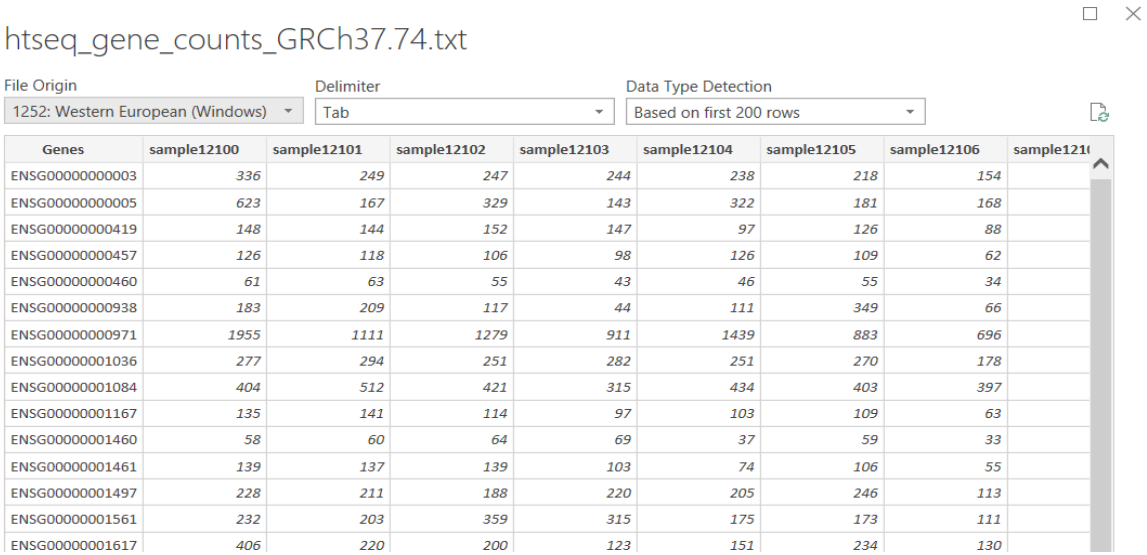
5.3 HTSEQ_Counts

24 Columns and 63684 rows - provides patients gene counts

Table 5 HTSEQ_Counts Columns

Column1	Genes	Gene ID from Ensembl Database

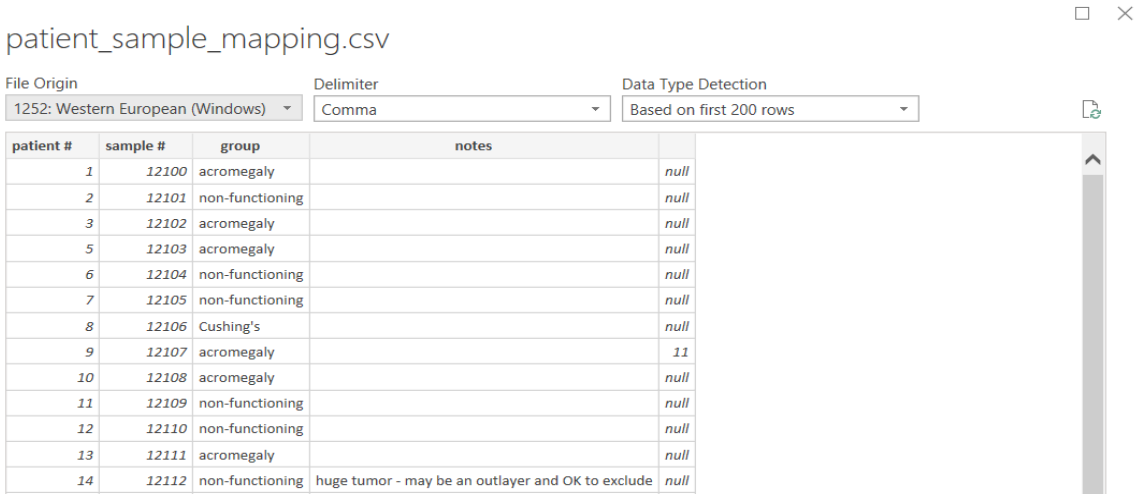
Column2 to 24	Sample121xx	Gene counts for 23 patients respectively



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

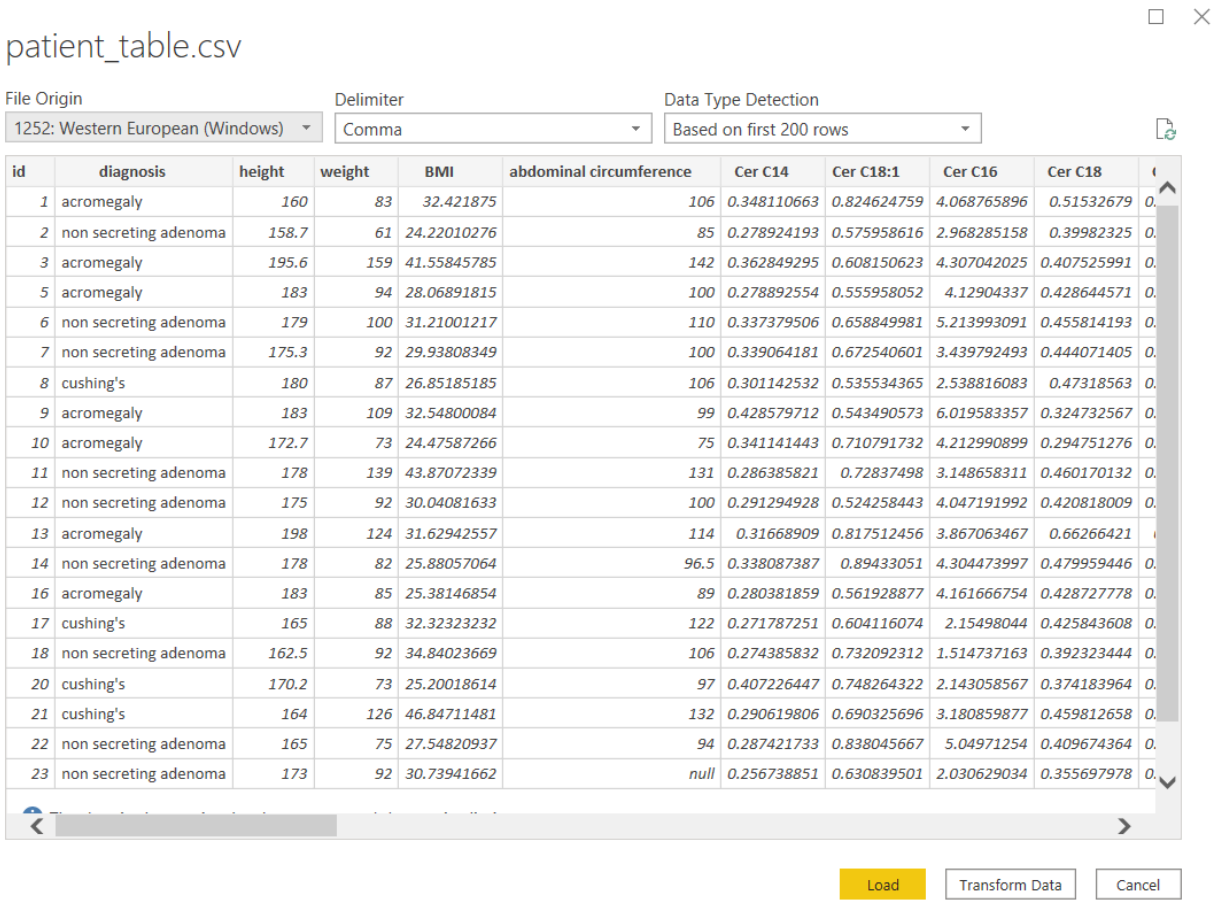
column 4 and column5 have no useful information for analysis.

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



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
2	non secreting adenoma	158.7	61	24.22010276	85	0.278924193	0.575958616	2.968285158	0.39982325
3	acromegaly	195.6	159	41.55845785	142	0.362849295	0.608150623	4.307042025	0.407525991
5	acromegaly	183	94	28.06891815	100	0.278892554	0.555958052	4.12904337	0.428644571
6	non secreting adenoma	179	100	31.21001217	110	0.337379506	0.658849981	5.213993091	0.455814193
7	non secreting adenoma	175.3	92	29.93808349	100	0.339064181	0.672540601	3.439792493	0.444071405
8	cushing's	180	87	26.85185185	106	0.301142532	0.535534365	2.538816083	0.47318563
9	acromegaly	183	109	32.54800084	99	0.428579712	0.543490573	6.019583357	0.324732567
10	acromegaly	172.7	73	24.47587266	75	0.341141443	0.710791732	4.212990899	0.294751276
11	non secreting adenoma	178	139	43.87072339	131	0.286385821	0.72837498	3.148658311	0.460170132
12	non secreting adenoma	175	92	30.04081633	100	0.291294928	0.524258443	4.047191992	0.420818009
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
16	acromegaly	183	85	25.38146854	89	0.280381859	0.561928877	4.161666754	0.428727778
17	cushing's	165	88	32.32323232	122	0.271787251	0.604116074	2.15498044	0.425843608
18	non secreting adenoma	162.5	92	34.84023669	106	0.274385832	0.732092312	1.514737163	0.392323444
20	cushing's	170.2	73	25.20018614	97	0.407226447	0.748264322	2.143058567	0.374183964
21	cushing's	164	126	46.84711481	132	0.290619806	0.690325696	3.180859877	0.459812658
22	non secreting adenoma	165	75	27.54820937	94	0.287421733	0.838045667	5.04971254	0.409674364
23	non secreting adenoma	173	92	30.73941662	null	0.256738851	0.630839501	2.030629034	0.355697978

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	

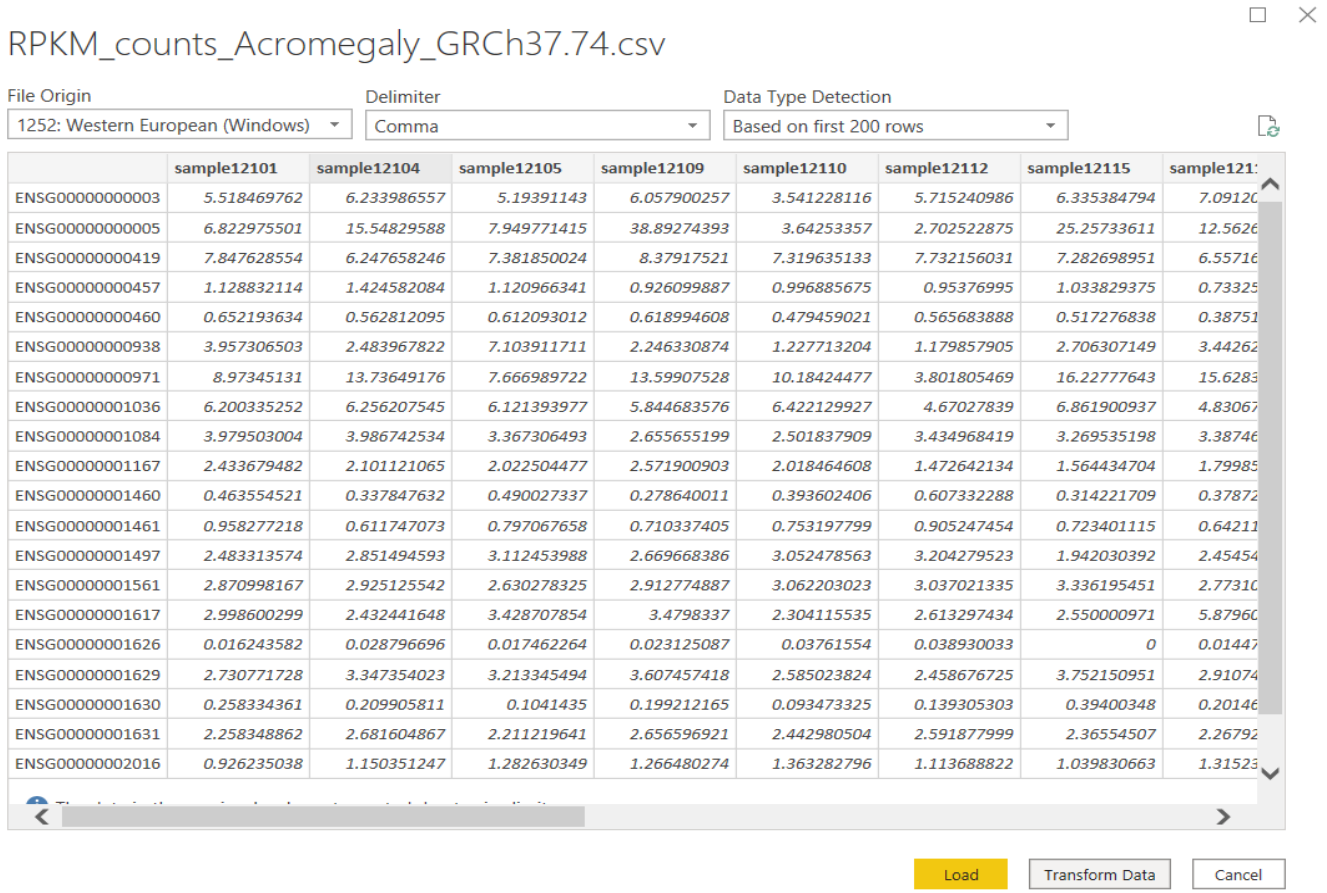
Column 31	AST	
Column 32	ALT	
Column 33	alk phos	
Column 34	HTN	
Column 35	diabetes	If the patient is diabetic or not
Column 36	smoking	Does the patient smoke or not

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

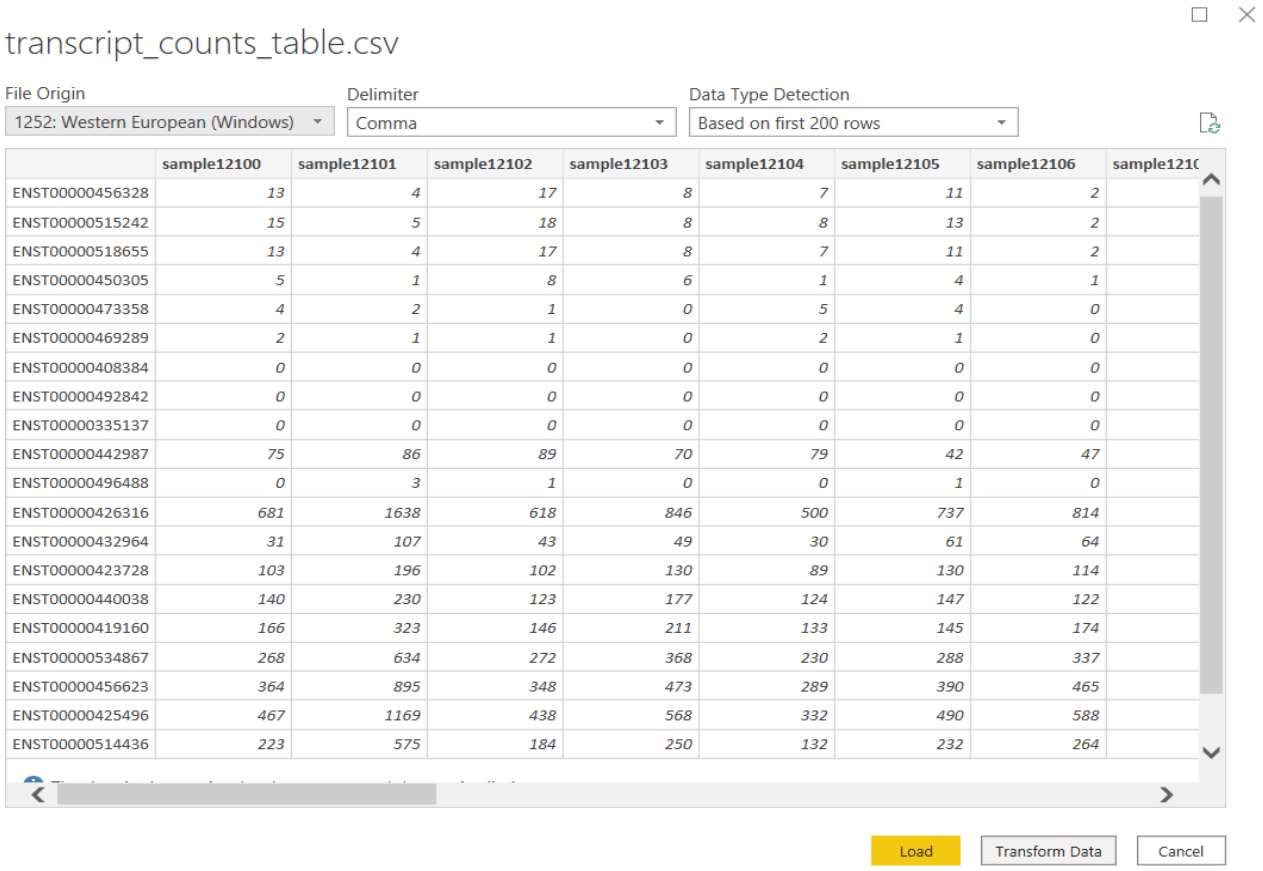


	sample12101	sample12104	sample12105	sample12109	sample12110	sample12112	sample12115	sample121
ENSG00000000003	5.518469762	6.233986557	5.19391143	6.057900257	3.541228116	5.715240986	6.335384794	7.0912
ENSG00000000005	6.822975501	15.54829588	7.949771415	38.89274393	3.64253357	2.702522875	25.25733611	12.562
ENSG00000000419	7.847628554	6.247658246	7.381850024	8.37917521	7.319635133	7.732156031	7.282698951	6.5571
ENSG00000000457	1.128832114	1.424582084	1.120966341	0.926099887	0.996885675	0.95376995	1.033829375	0.7332
ENSG00000000460	0.652193634	0.562812095	0.612093012	0.618994608	0.479459021	0.565683888	0.517276838	0.3875
ENSG00000000938	3.957306503	2.483967822	7.103911711	2.246330874	1.227713204	1.179857905	2.706307149	3.4426
ENSG00000000971	8.97345131	13.73649176	7.666989722	13.59907528	10.18424477	3.801805469	16.22777643	15.628
ENSG00000001036	6.200335252	6.256207545	6.121393977	5.844683576	6.422129927	4.67027839	6.861900937	4.8306
ENSG00000001084	3.979503004	3.986742534	3.367306493	2.655655199	2.501837909	3.434968419	3.269535198	3.3874
ENSG00000001167	2.433679482	2.101121065	2.022504477	2.571900903	2.018464608	1.472642134	1.564434704	1.7998
ENSG00000001460	0.463554521	0.337847632	0.490027337	0.278640011	0.393602406	0.607332288	0.314221709	0.3787
ENSG00000001461	0.958277218	0.611747073	0.797067658	0.710337405	0.753197799	0.905247454	0.723401115	0.6421
ENSG00000001497	2.483313574	2.851494593	3.112453988	2.669668386	3.052478563	3.204279523	1.942030392	2.4545
ENSG00000001561	2.870998167	2.925125542	2.630278325	2.912774887	3.062203023	3.037021335	3.336195451	2.7731
ENSG00000001617	2.998600299	2.432441648	3.428707854	3.4798337	2.304115535	2.613297434	2.550000971	5.8796
ENSG00000001626	0.016243582	0.028796696	0.017462264	0.023125087	0.03761554	0.038930033	0	0.0144
ENSG00000001629	2.730771728	3.347354023	3.213345494	3.607457418	2.585023824	2.458676725	3.752150951	2.9107
ENSG00000001630	0.258334361	0.209905811	0.1041435	0.199212165	0.093473325	0.139305303	0.39400348	0.2014
ENSG00000001631	2.258348862	2.681604867	2.211219641	2.656596921	2.442980504	2.591877999	2.36554507	2.2679
ENSG00000002016	0.926235038	1.150351247	1.282630349	1.266480274	1.363282796	1.113688822	1.039830663	1.3152

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



	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	

B) Data Pre-Processing and Data Cleansing

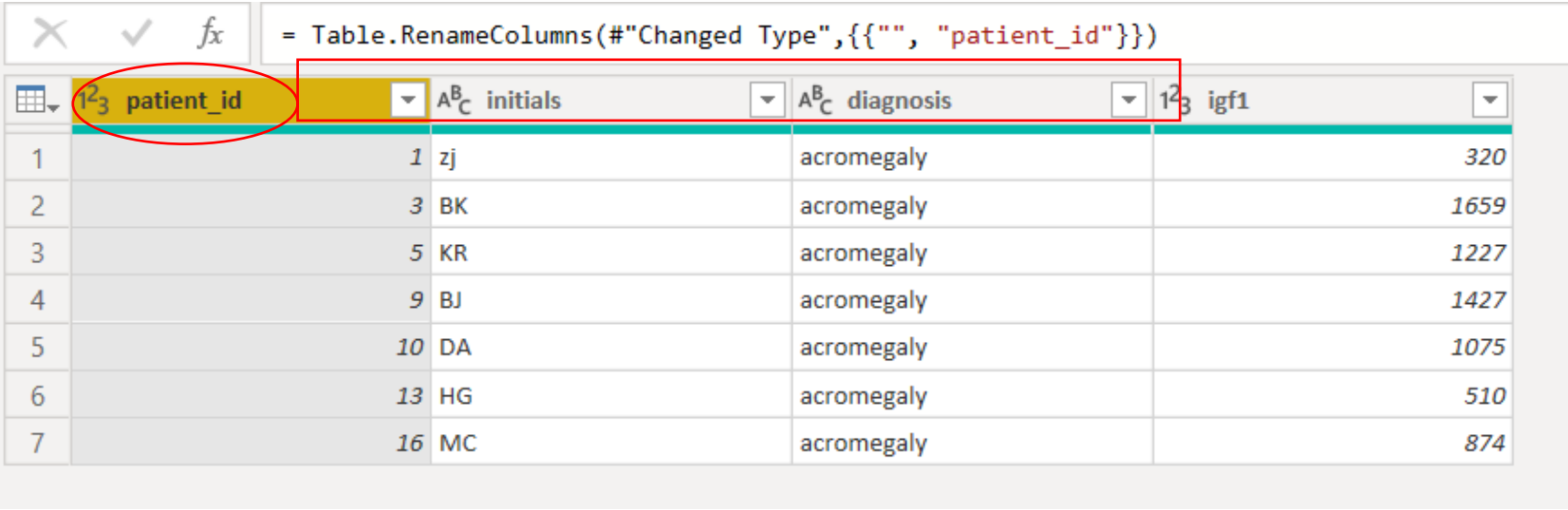
Is there any evidence of steps performed to cleanse the data? For example: • Removing NAs, • Renaming columns • Changing data types • Removing errors • Removing columns • Merging tables, etc

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"}})**

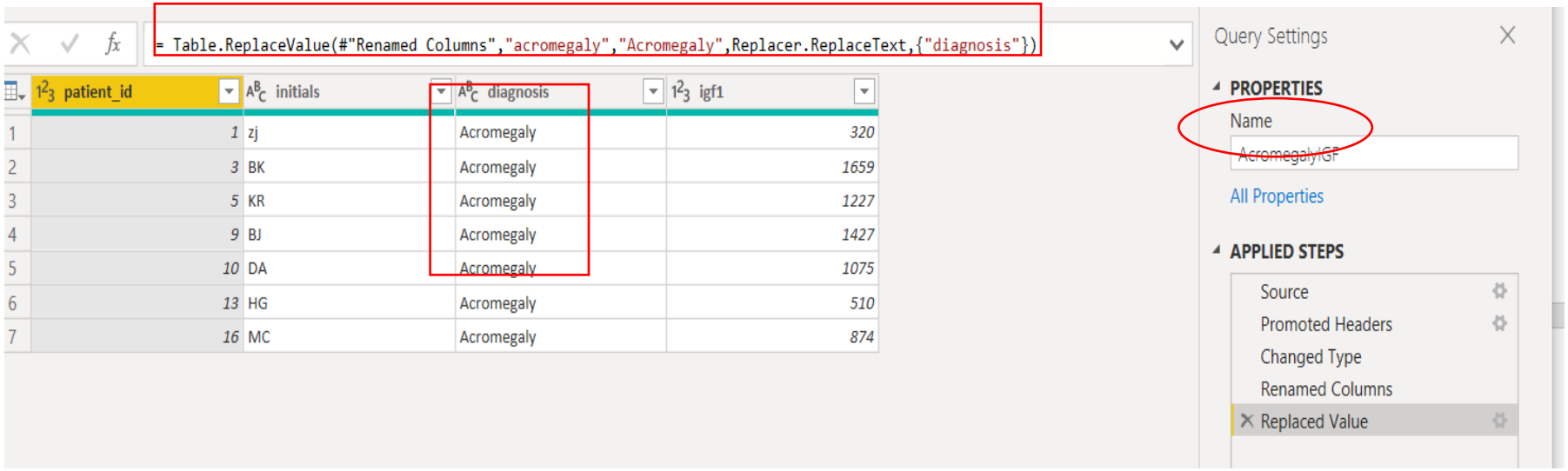


	patient_id	initials	diagnosis	igf1
1	1	zj	acromegaly	320
2	3	BK	acromegaly	1659
3	5	KR	acromegaly	1227
4	9	BJ	acromegaly	1427
5	10	DA	acromegaly	1075
6	13	HG	acromegaly	510
7	16	MC	acromegaly	874

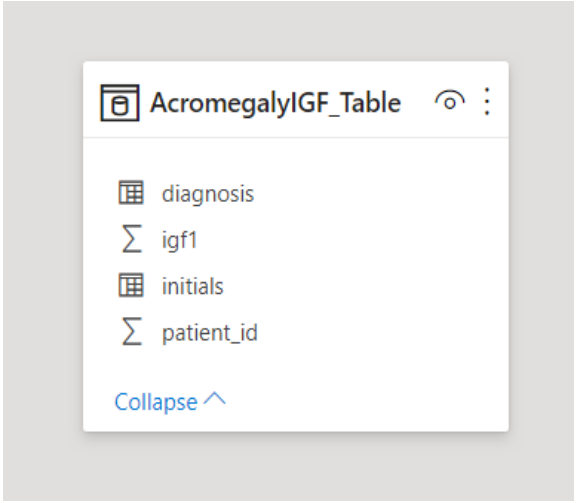
1.2. Replacing Values

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"})**



1.3. Model View



2. Table 2: Ensembl Gene Annotation

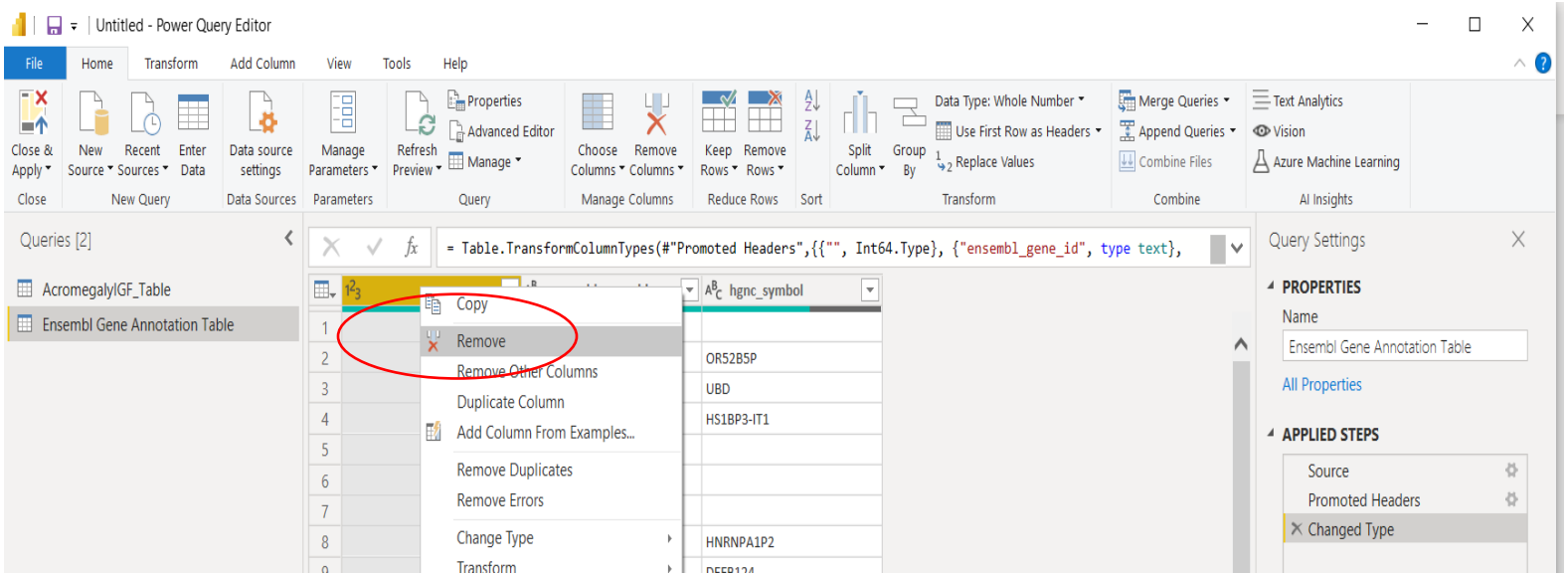
Ensembl Gene Annotation.csv

File Origin: 1252: Western European (Windows) | Delimiter: Comma | Data Type Detection: Based on first 200 ro

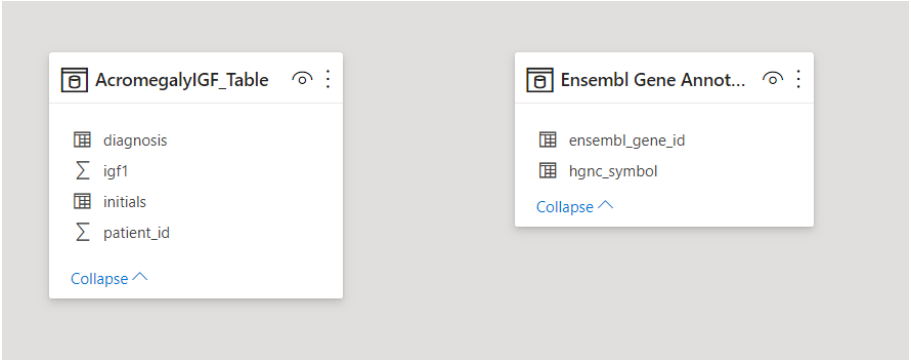
	ensembl_gene_id	hgnc_symbol
1	ENSG00000197468	
2	ENSG00000231049	OR52B5P
3	ENSG00000228913	UBD
4	ENSG00000231948	HS1BP3-IT1
5	ENSG00000231510	
6	ENSG00000229336	
7	ENSG00000261641	
8	ENSG00000237295	HNRNPA1P2
9	ENSG00000180383	DEFB124
10	ENSG00000229093	OR51AB1P
11	ENSG00000270100	
12	ENSG00000272894	

2.1. Removing Columns

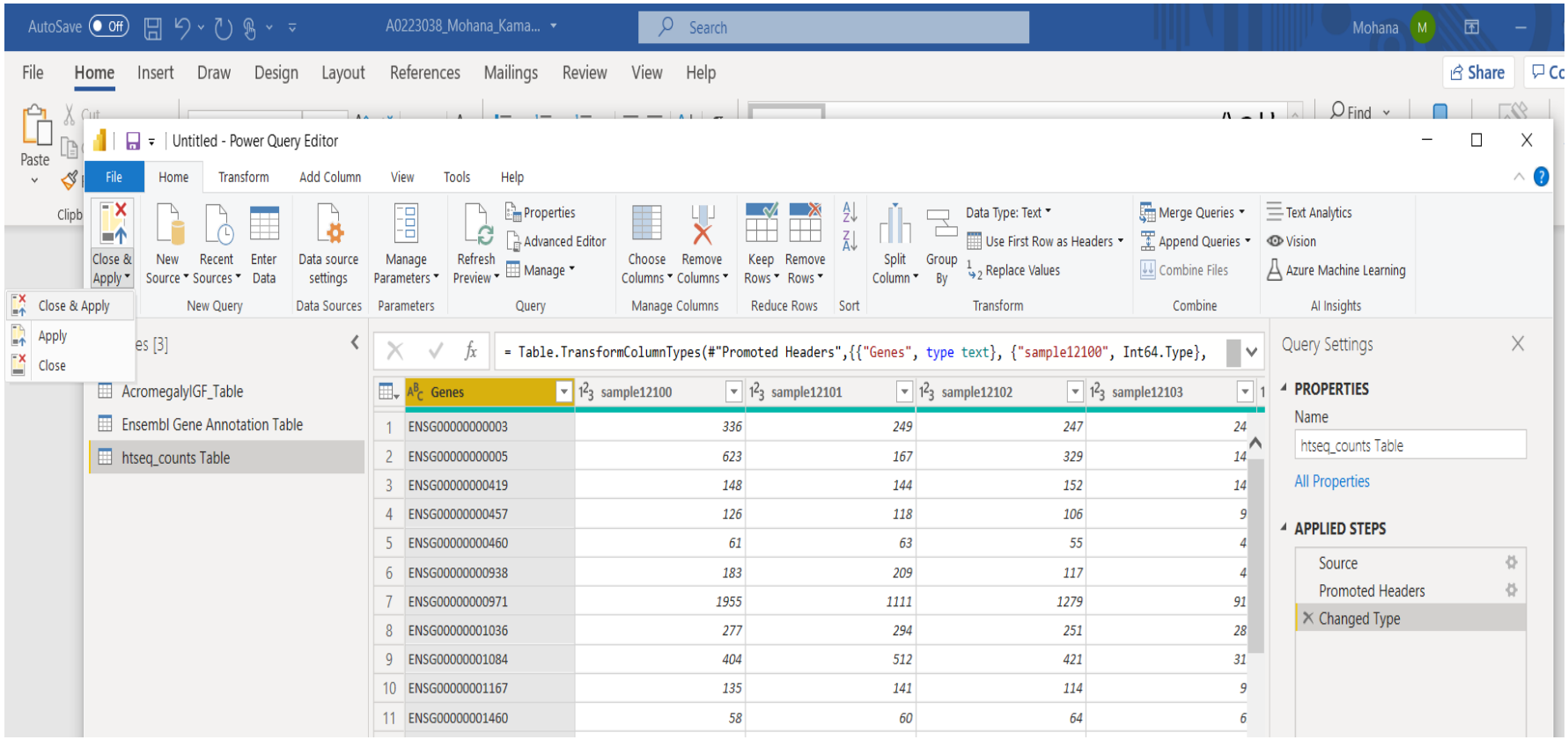
The first blank column has index values and is not very useful in analysis. Removing the column as below.



2.2. Model View

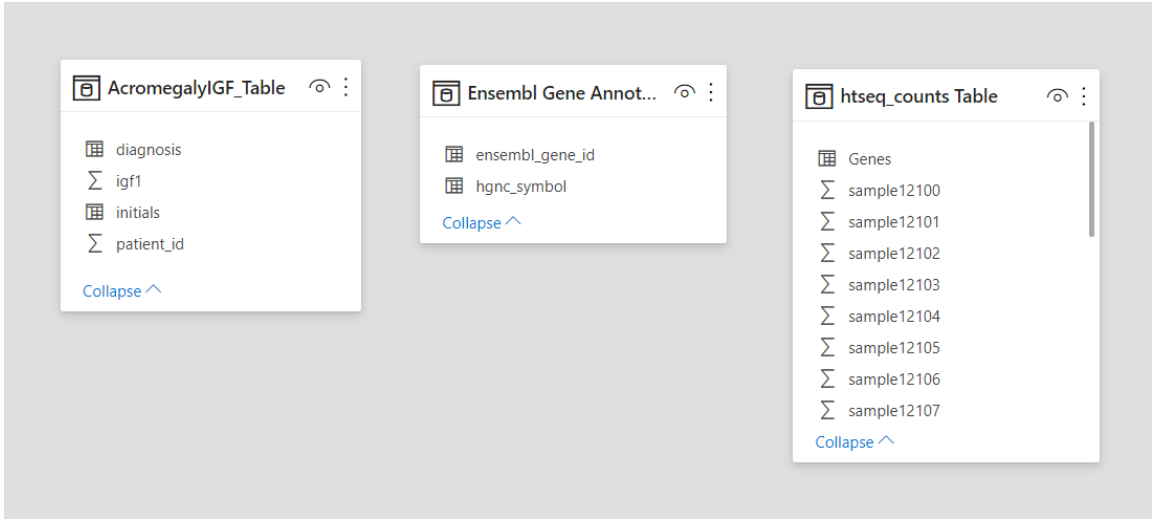


3. Table 3: HTSEQ_Counts

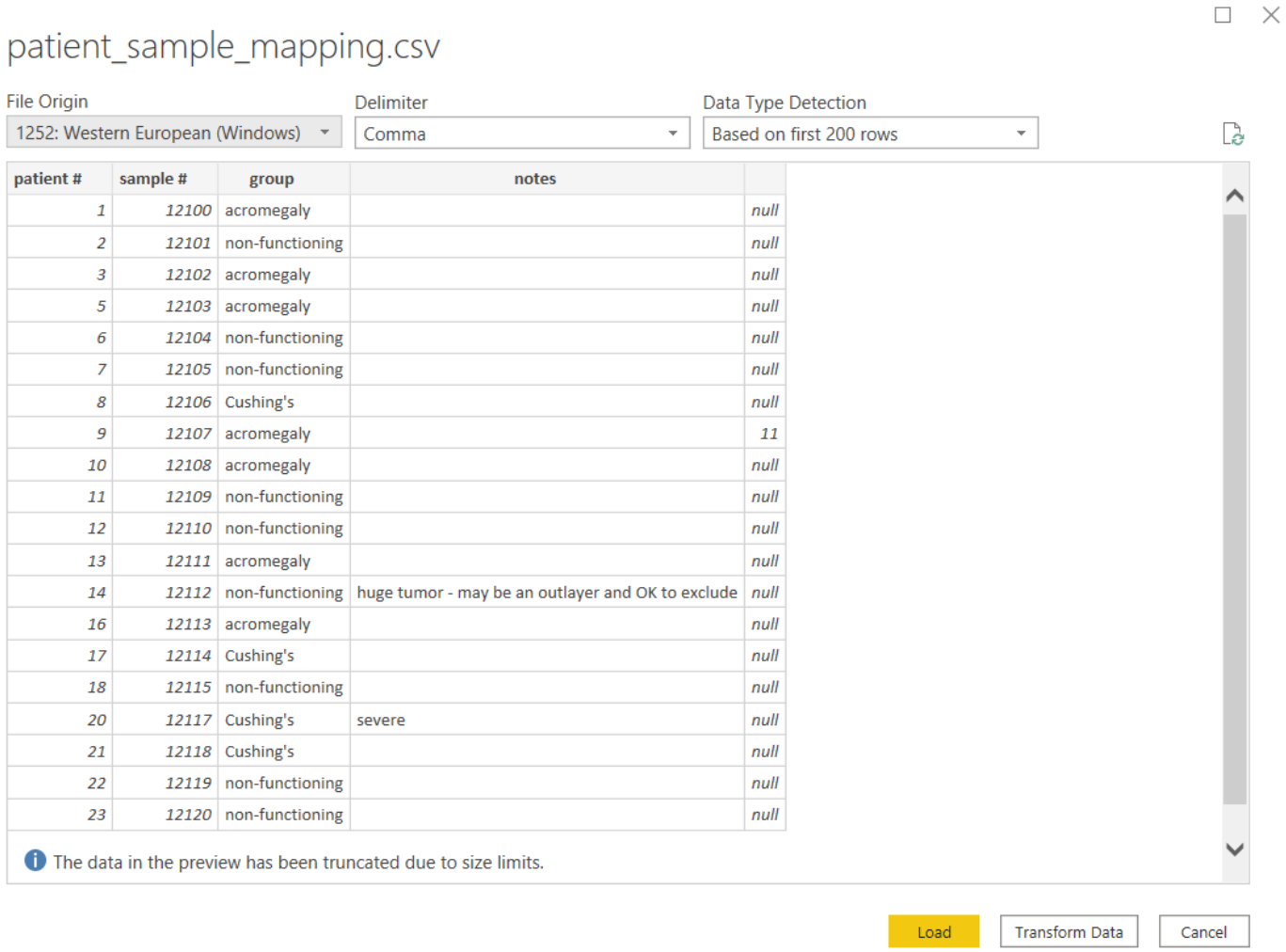


	Genes	sample12100	sample12101	sample12102	sample12103
1	ENSG00000000003	336	249	247	24
2	ENSG00000000005	623	167	329	14
3	ENSG00000000419	148	144	152	14
4	ENSG00000000457	126	118	106	9
5	ENSG00000000460	61	63	55	4
6	ENSG00000000938	183	209	117	4
7	ENSG00000000971	1955	1111	1279	91
8	ENSG00000001036	277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

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.

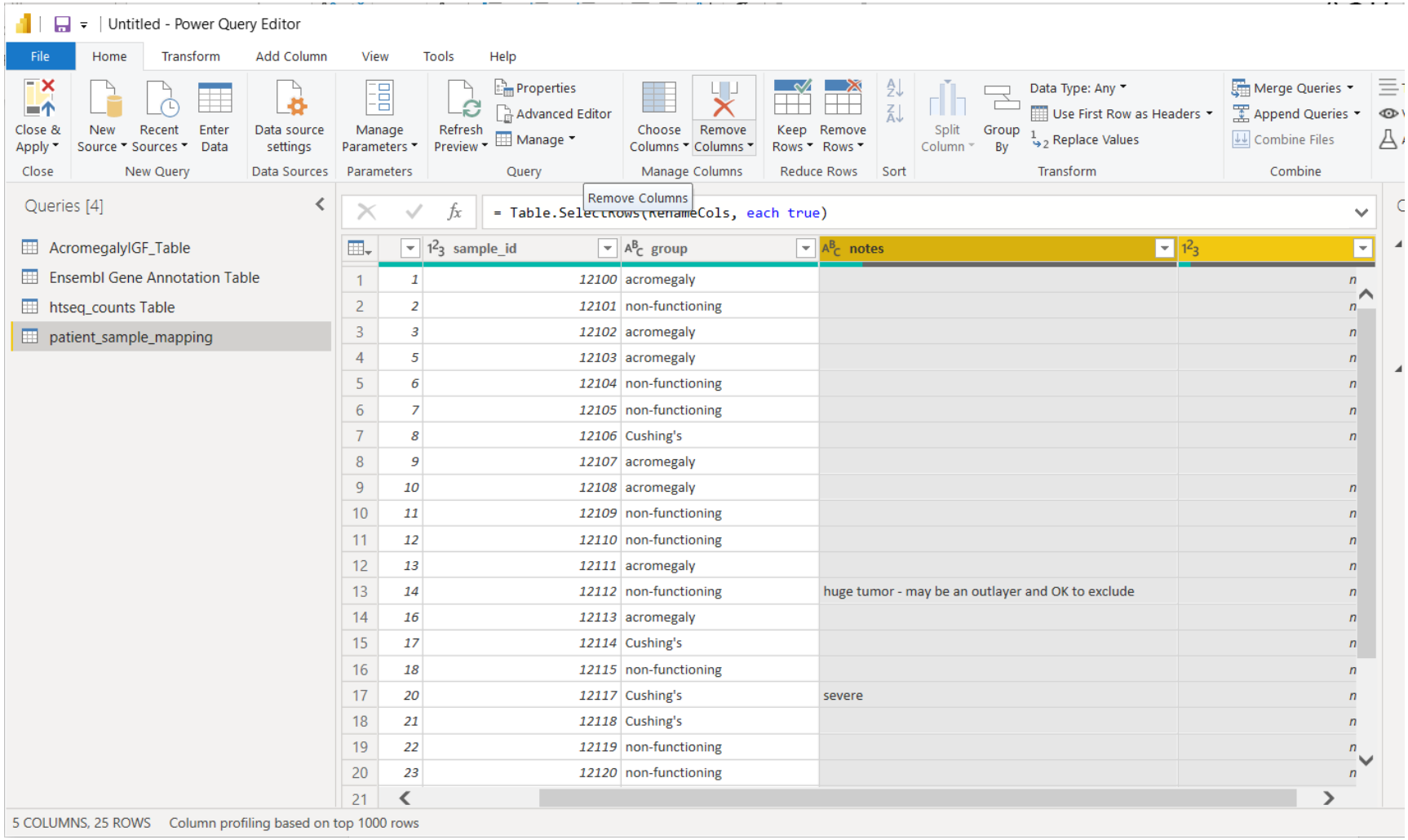
Load | Transform Data | Cancel

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.

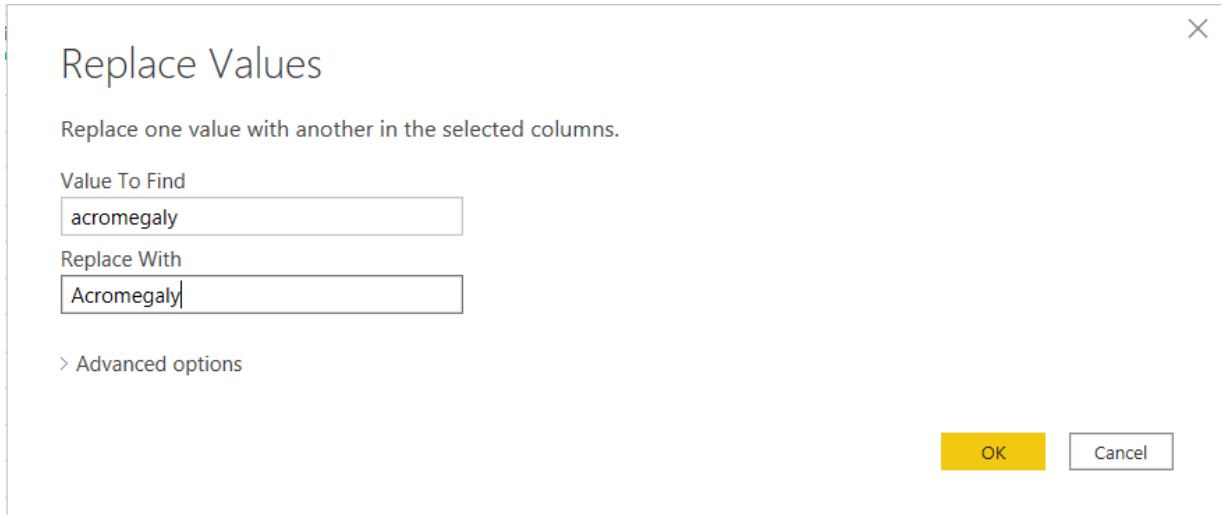
4.2. Removing Columns

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.

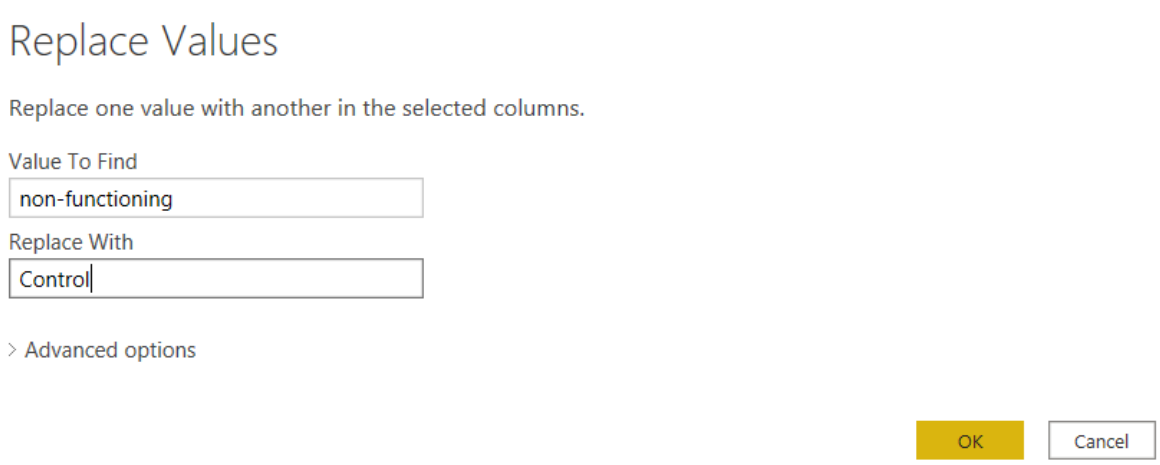


4.3. Replacing Values

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

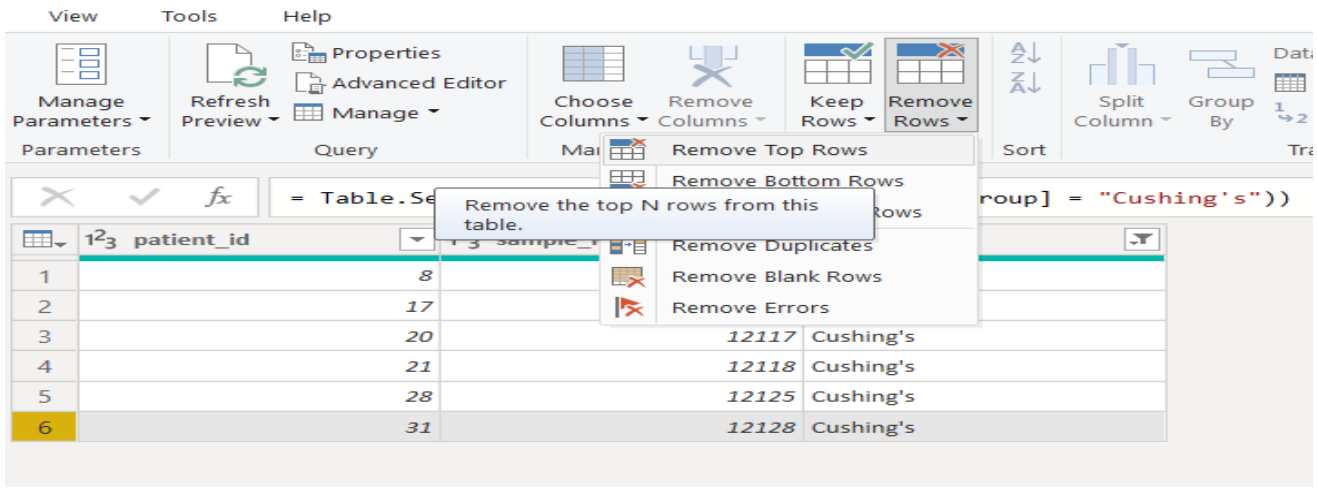


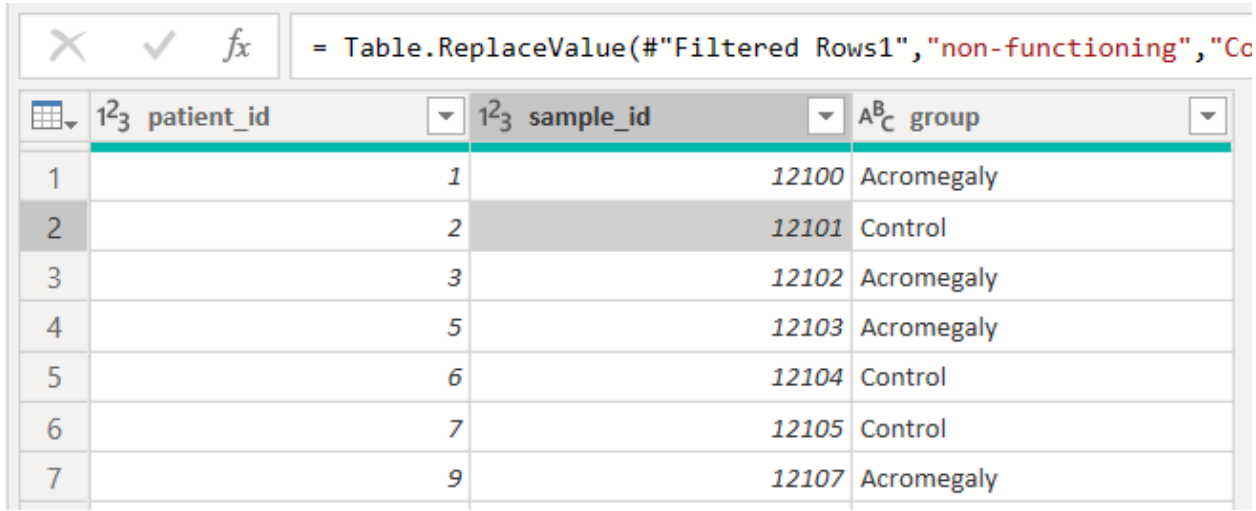
Also, replace “non-functioning” with Control



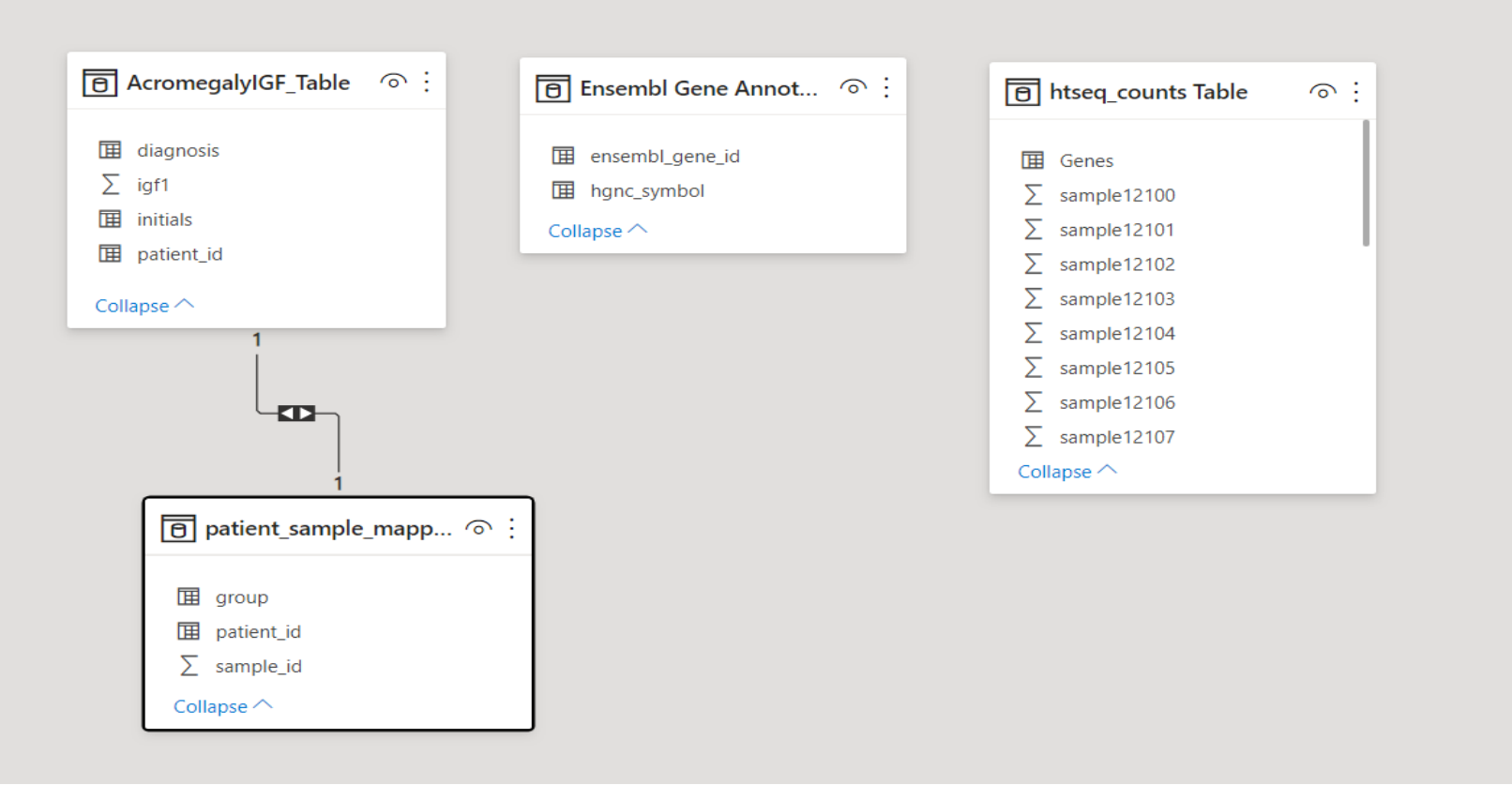
4.4. Filtering Rows

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.

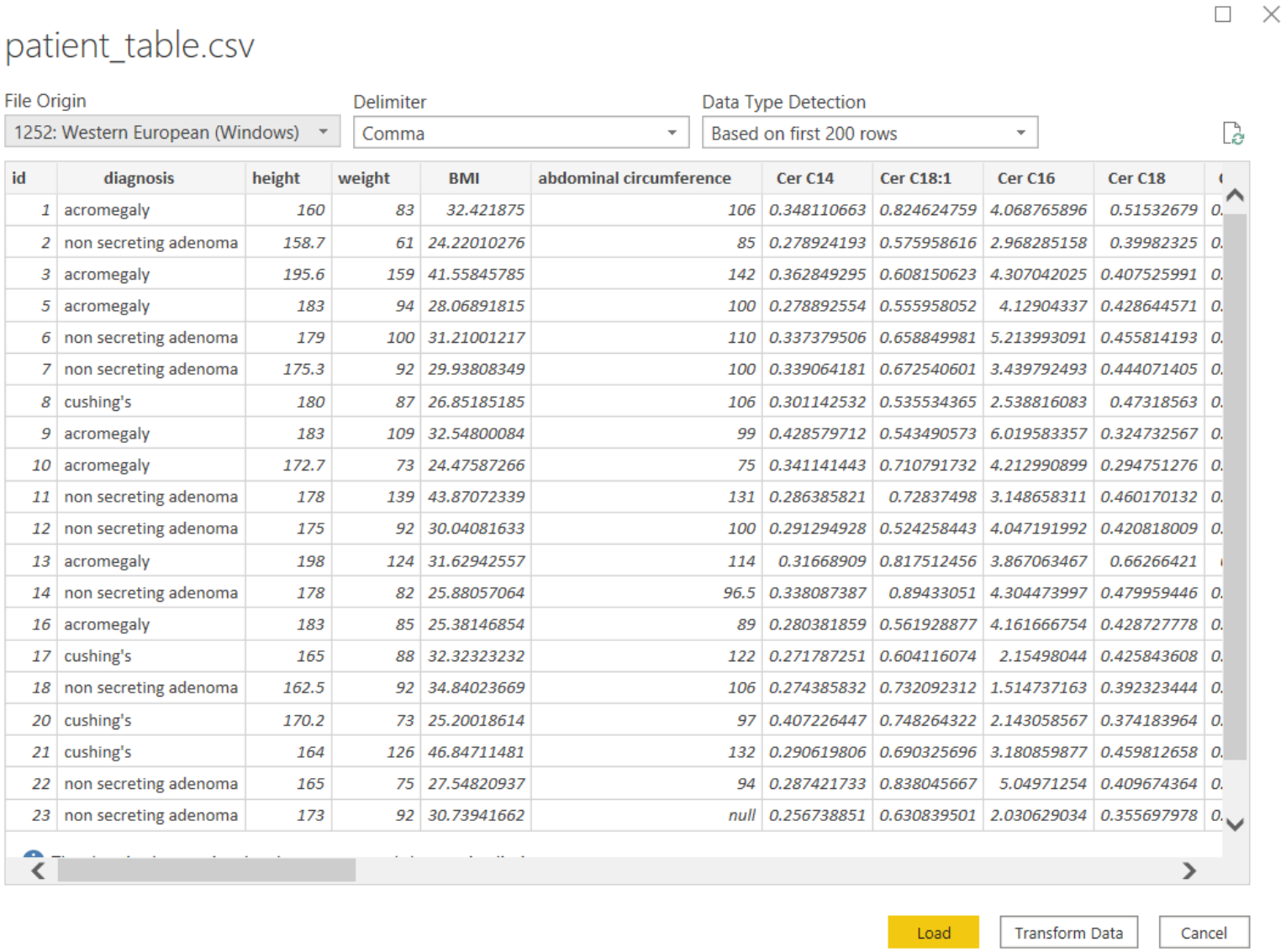

View
Tools
Help
Manage Parameters
Parameters
Refresh Preview
Properties
Advanced Editor
Manage
Query
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Remove Top Rows
Remove Bottom Rows
Remove Duplicates
Remove Blank Rows
Remove Errors
Remove the top N rows from this table.
Sort
Split Column
Group By
= Table.Se
roup] = "Cushing's"))
patient_id
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

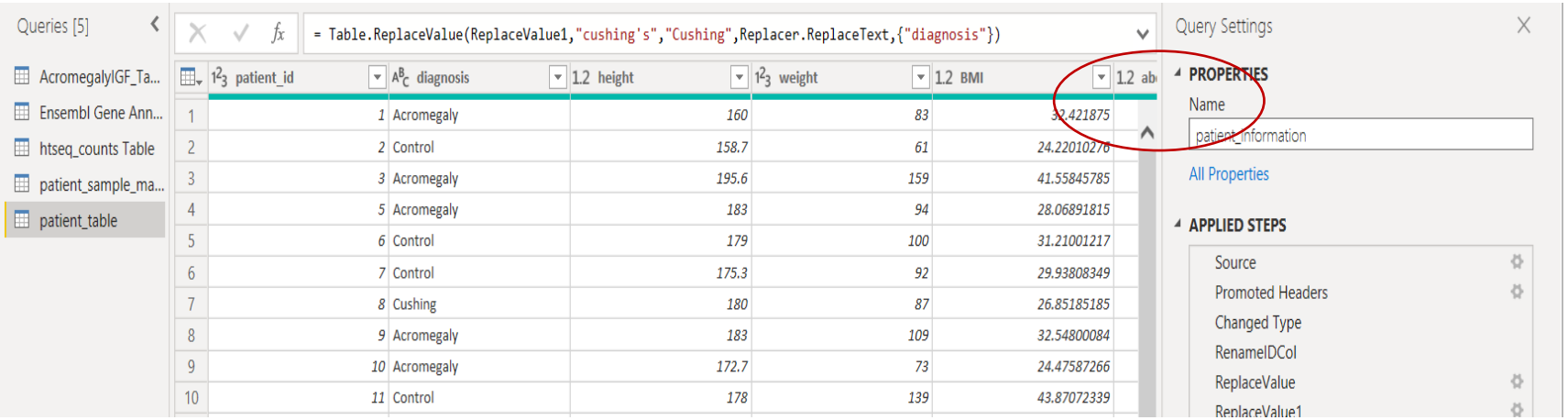
File Origin: 1252: Western European (Windows) | Delimiter: Comma | Data Type Detection: Based on first 200 rows

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.

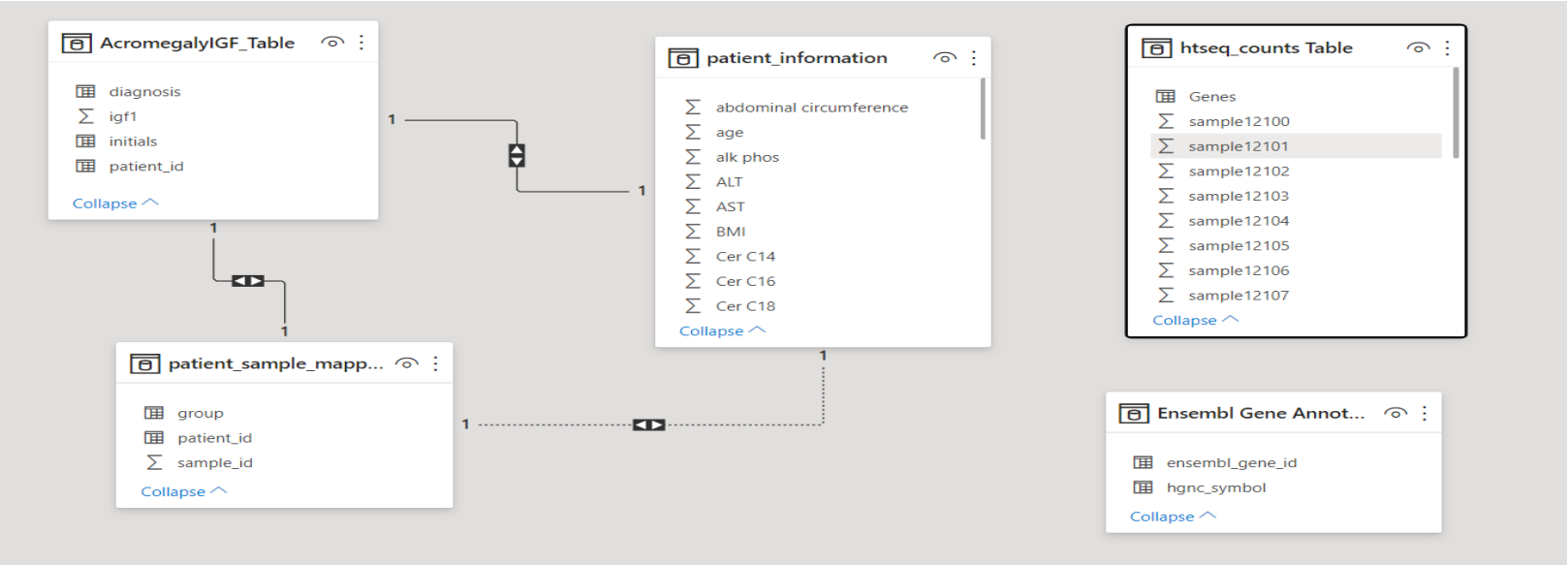
Load | Transform Data | Cancel

5.1. Replace Values

5.2. Renaming the table



5.3. Model View



6. Table 6: IGF_RKPM_Count

6.1. Rename Column

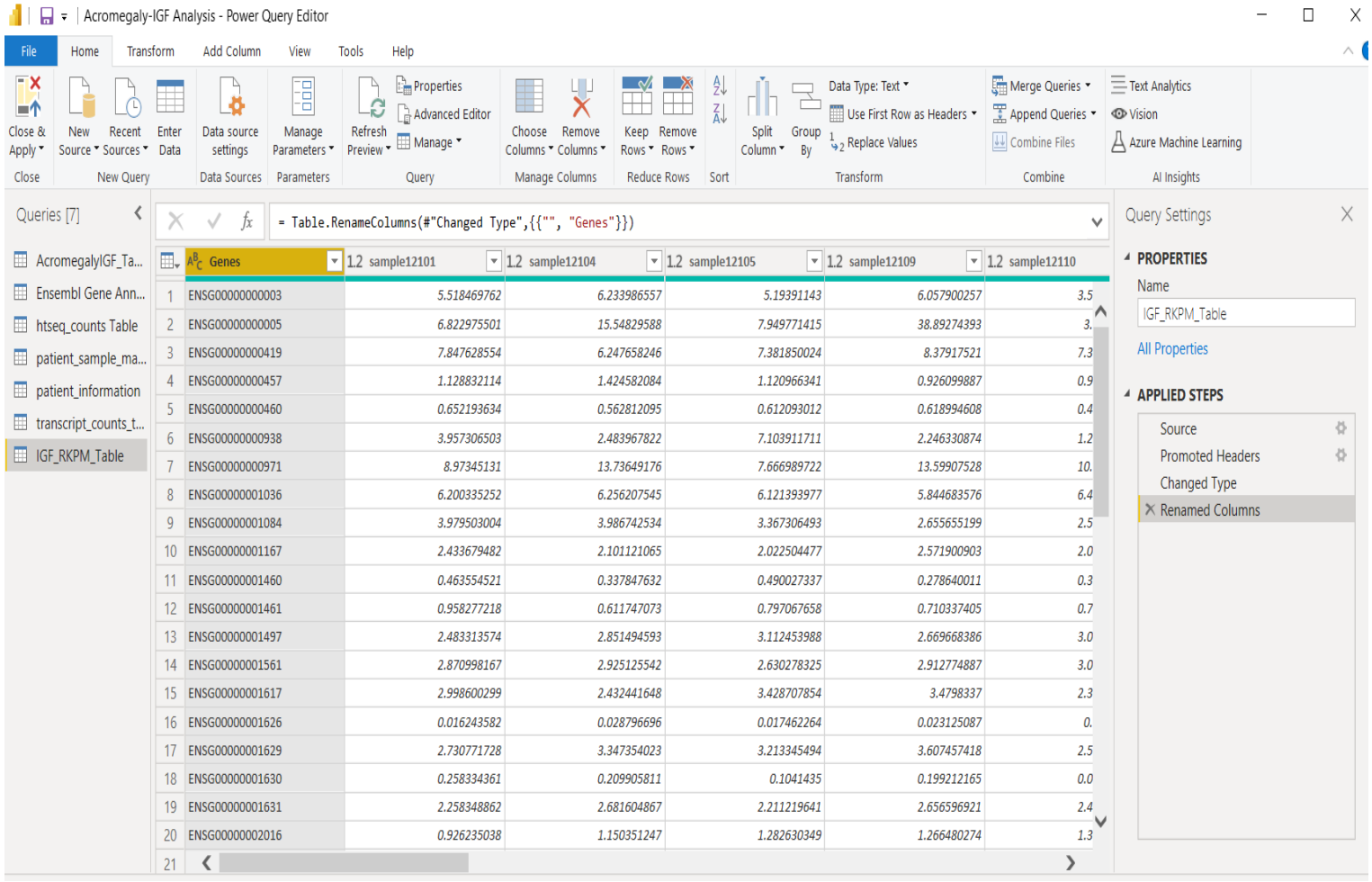
Import the table as and rename the first column to “Genes”

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

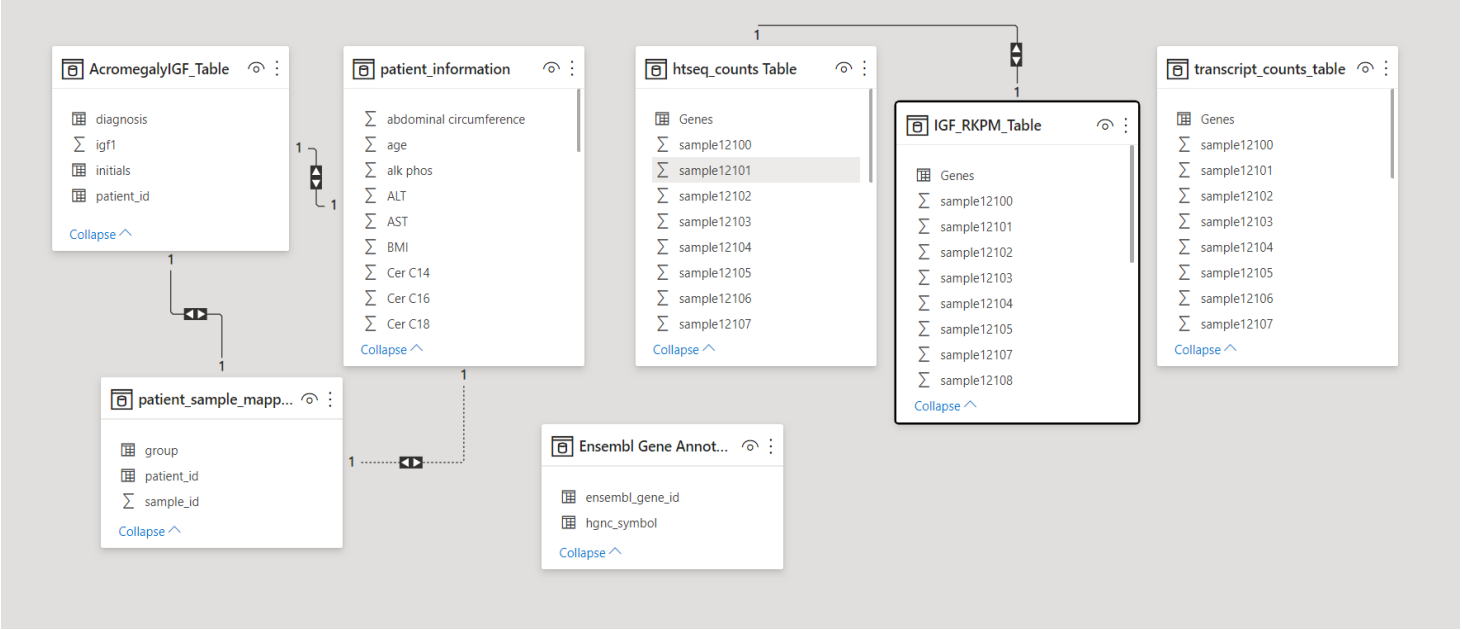
RPKM_counts_Acromegaly_GRCh37.74.csv

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	7.0912
ENSG00000000005	6.822975501	15.54829588	7.949771415	38.89274393	3.64253357	2.702522875	25.25733611	12.562
ENSG00000000419	7.847628554	6.247658246	7.381850024	8.37917521	7.319635133	7.732156031	7.282698951	6.5571
ENSG00000000457	1.128832114	1.424582084	1.120966341	0.926099887	0.996885675	0.95376995	1.033829375	0.7332
ENSG00000000460	0.652193634	0.562812095	0.612093012	0.618994608	0.479459021	0.565683888	0.517276838	0.3875
ENSG00000000938	3.957306503	2.483967822	7.103911711	2.246330874	1.227713204	1.179857905	2.706307149	3.4426
ENSG00000000971	8.97345131	13.73649176	7.666989722	13.59907528	10.18424477	3.801805469	16.22777643	15.628
ENSG00000001036	6.200335252	6.256207545	6.121393977	5.844683576	6.422129927	4.67027839	6.861900937	4.8306
ENSG00000001084	3.979503004	3.986742534	3.367306493	2.655655199	2.501837909	3.434968419	3.269535198	3.3874
ENSG00000001167	2.433679482	2.101121065	2.022504477	2.571900903	2.018464608	1.472642134	1.564434704	1.7998
ENSG00000001460	0.463554521	0.337847632	0.490027337	0.278640011	0.393602406	0.607332288	0.314221709	0.3787


Acromegaly-IGF Analysis - Power Query Editor
File
Home
Transform
Add Column
View
Tools
Help
Close & Apply
New Source
Recent Sources
Enter Data
Data source settings
Manage Parameters
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Merge Queries
Append Queries
Combine Files
Text Analytics
Vision
Azure Machine Learning
Close
New Query
Data Sources
Parameters
Query
Manage Columns
Reduce Rows
Sort
Transform
Combine
AI Insights
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

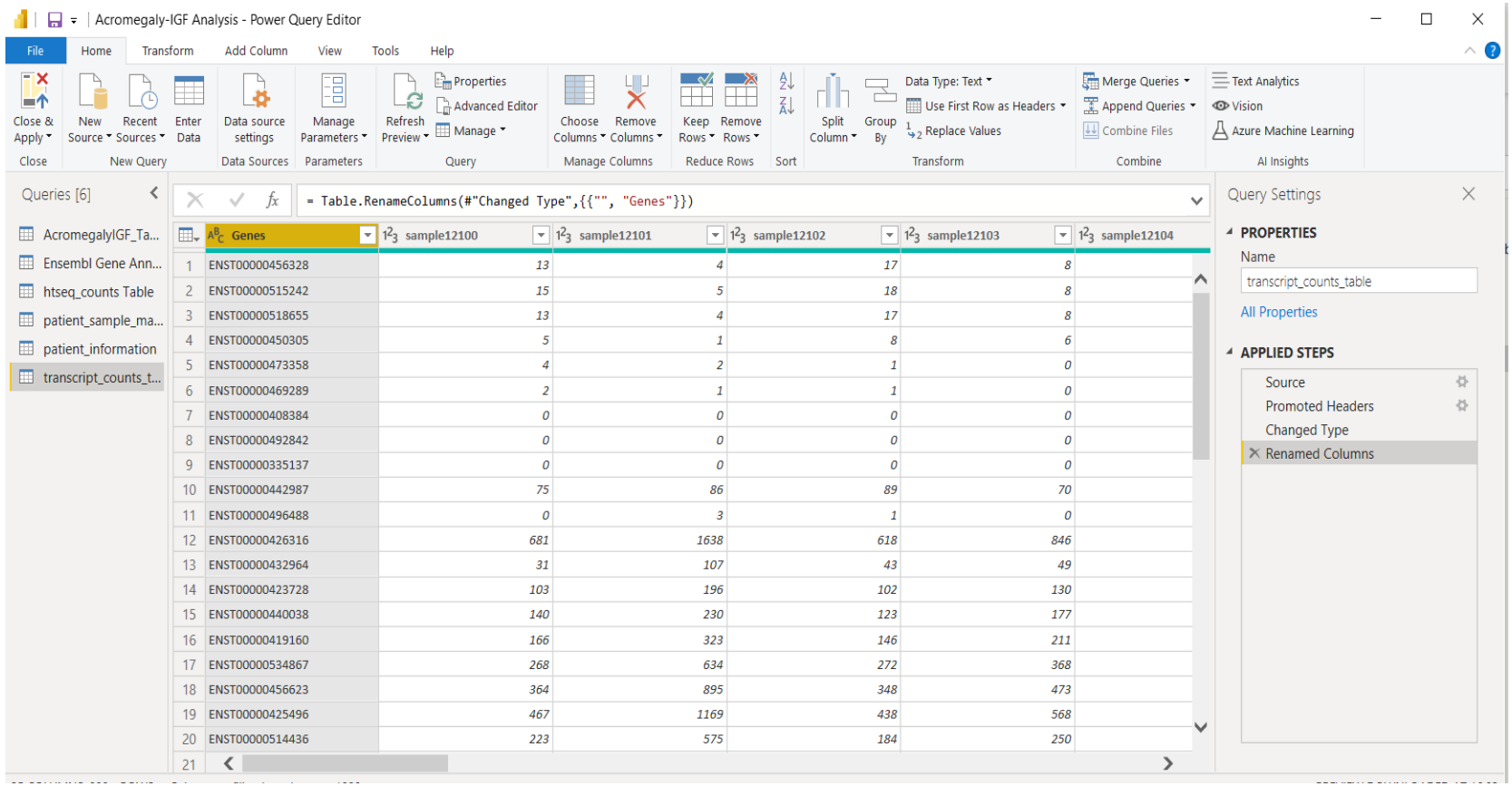
	sample12100	sample12101	sample12102	sample12103	sample12104	sample12105	sample12106	sample1210
ENST00000456328	13	4	17	8	7	11	2	
ENST00000515242	15	5	18	8	8	13	2	
ENST00000518655	13	4	17	8	7	11	2	
ENST00000450305	5	1	8	6	1	4	1	
ENST00000473358	4	2	1	0	5	4	0	
ENST00000469289	2	1	1	0	2	1	0	
ENST00000408384	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	

Load | Transform Data | Cancel

7.1. Rename Column

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

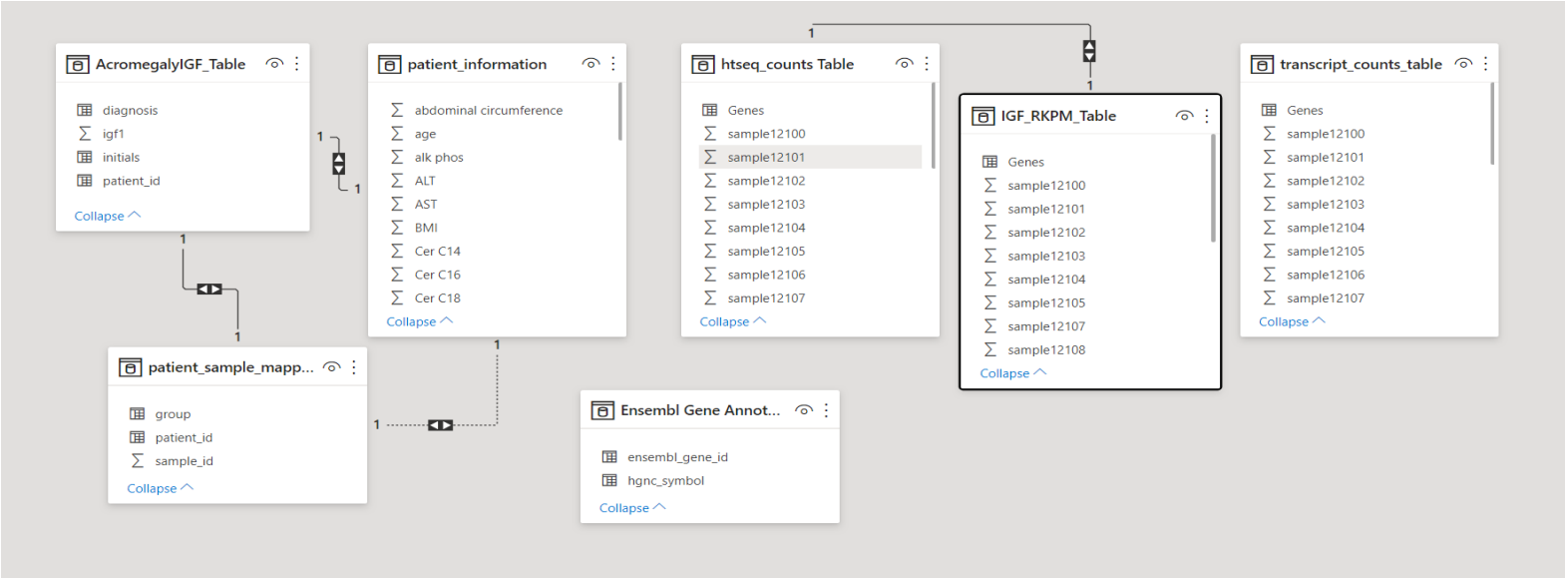
RenameCols = Table.RenameColumns(#"Changed Type",{{"", "Genes"}})



C) Data Modelling –Schema Facts and Dimensions

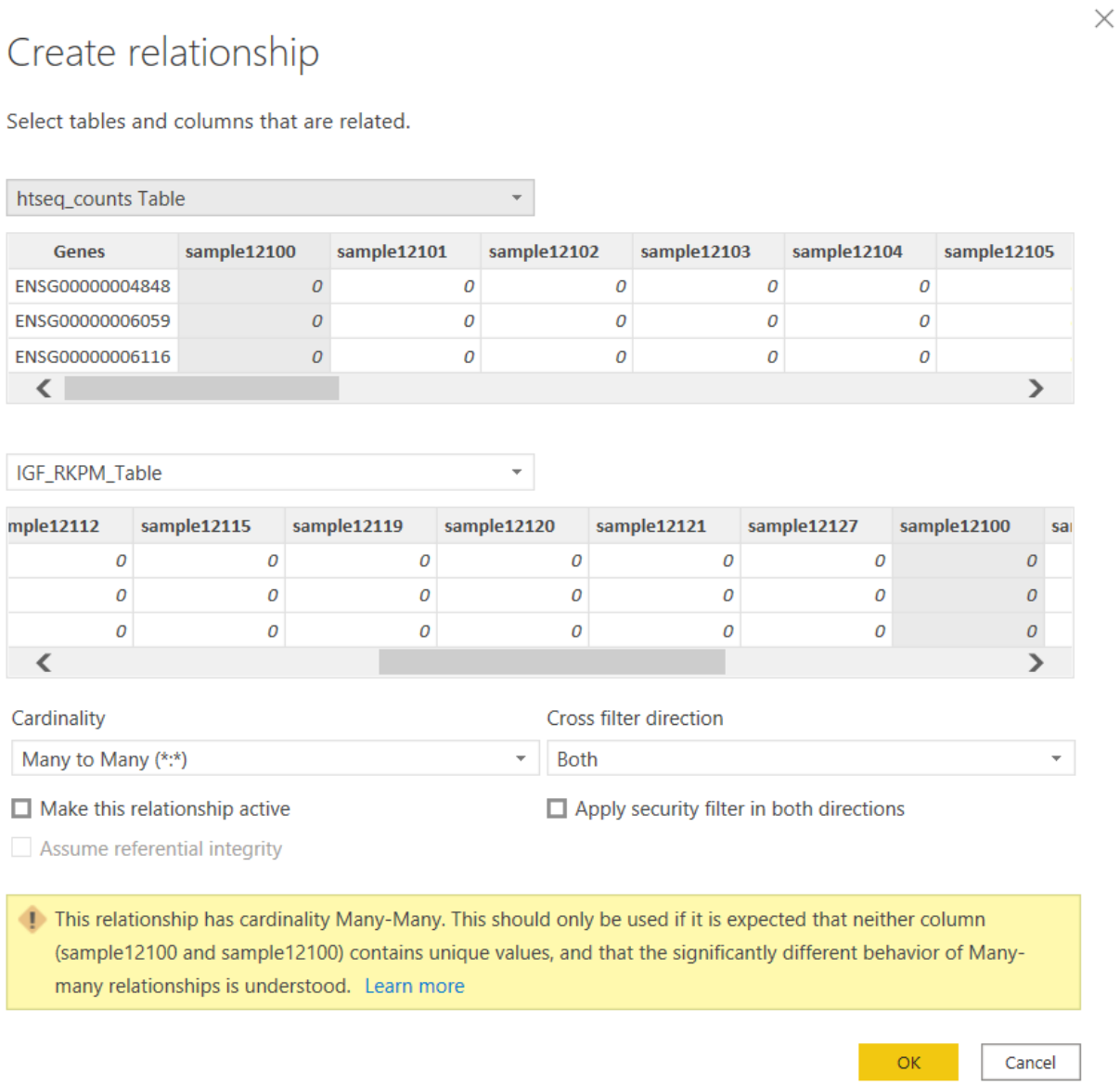
1. Data Modelling Process

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.

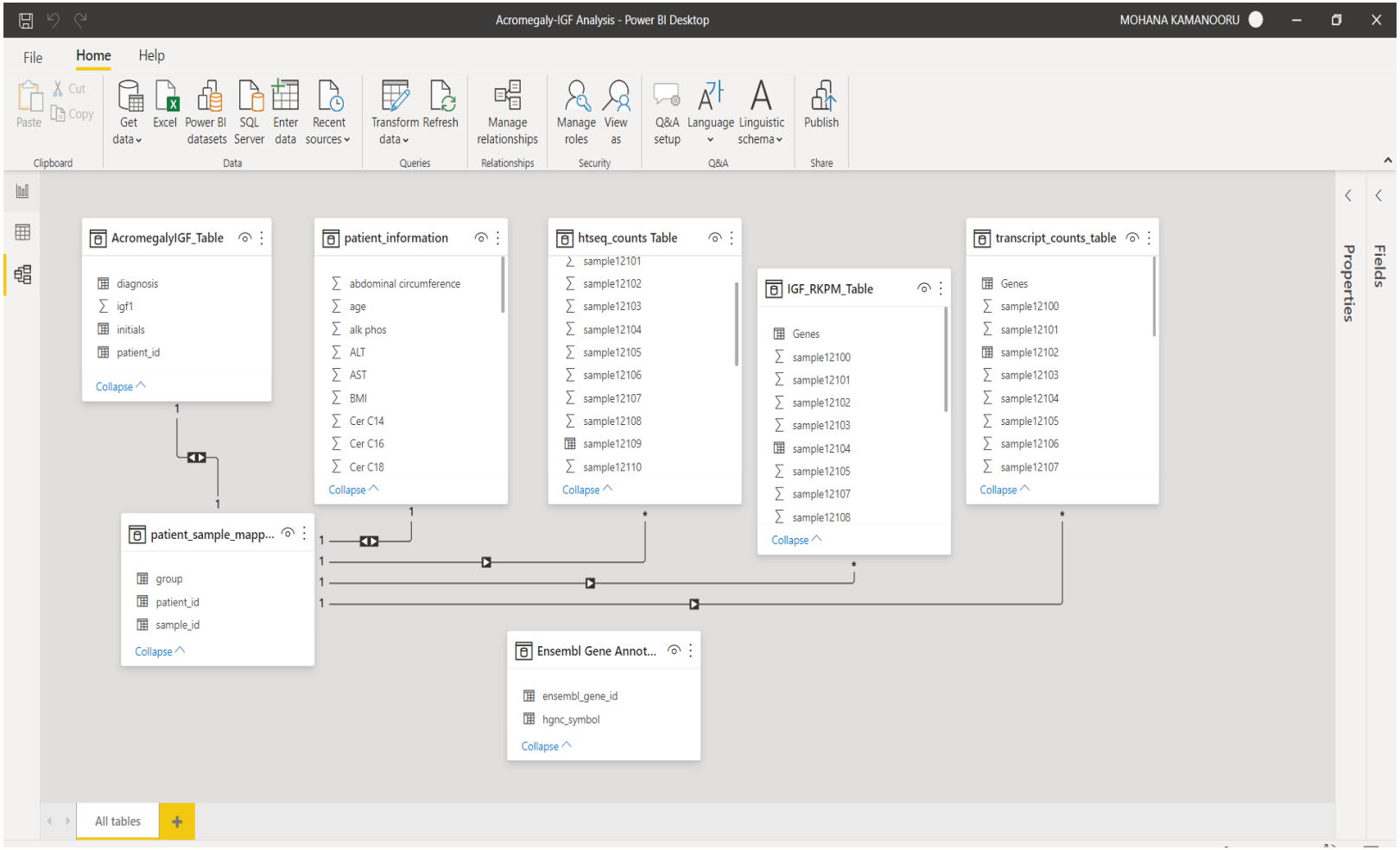


1.1. Creating Relationships

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.

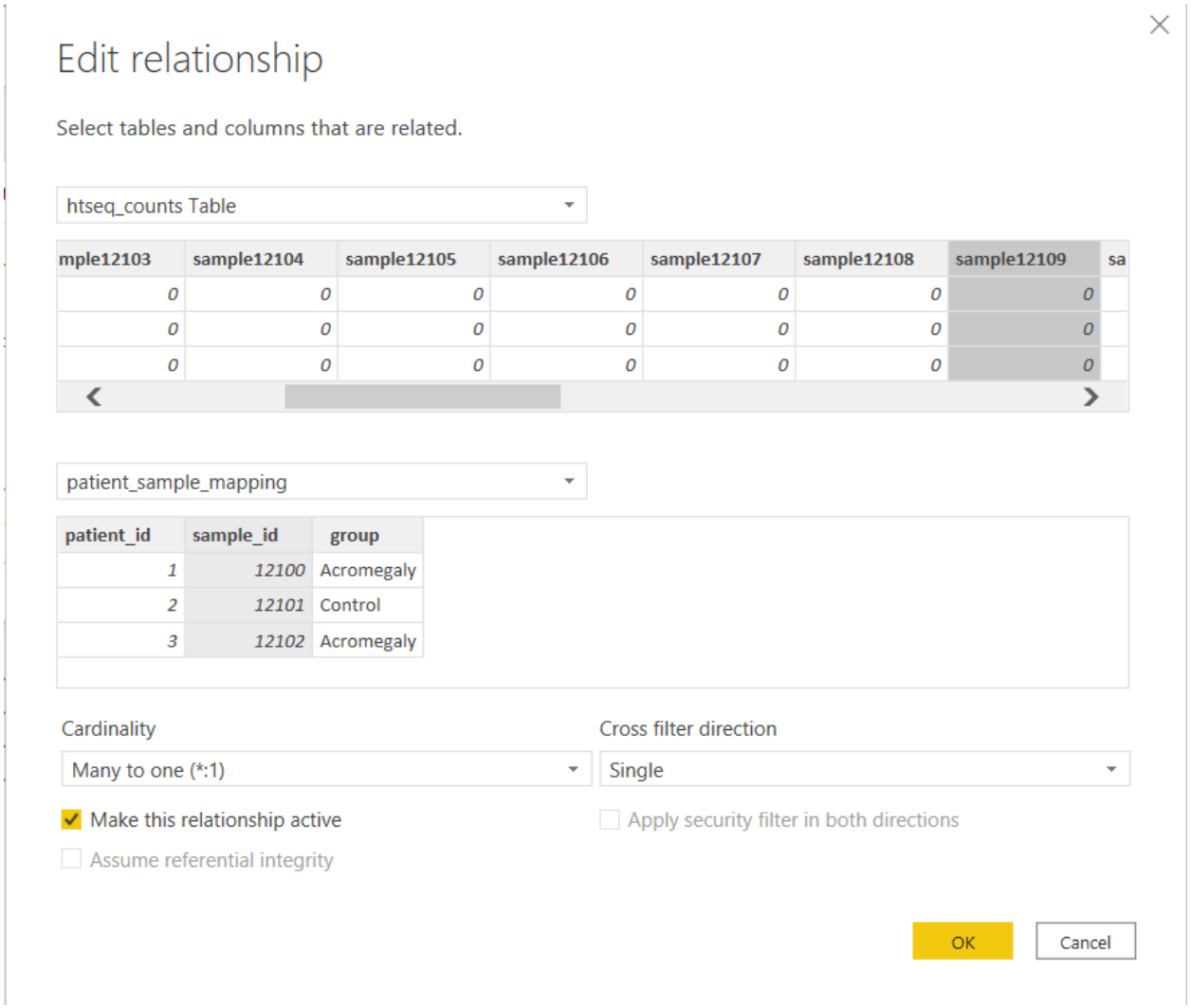

Create relationship
Select tables and columns that are related.
htseq_counts Table
Genes sample12100 sample12101 sample12102 sample12103 sample12104 sample12105
ENSG00000004848 0 0 0 0 0
ENSG00000006059 0 0 0 0 0
ENSG00000006116 0 0 0 0 0
IGF_RKPM_Table
mple12112 sample12115 sample12119 sample12120 sample12121 sample12127 sample12100
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
OK
Cancel

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



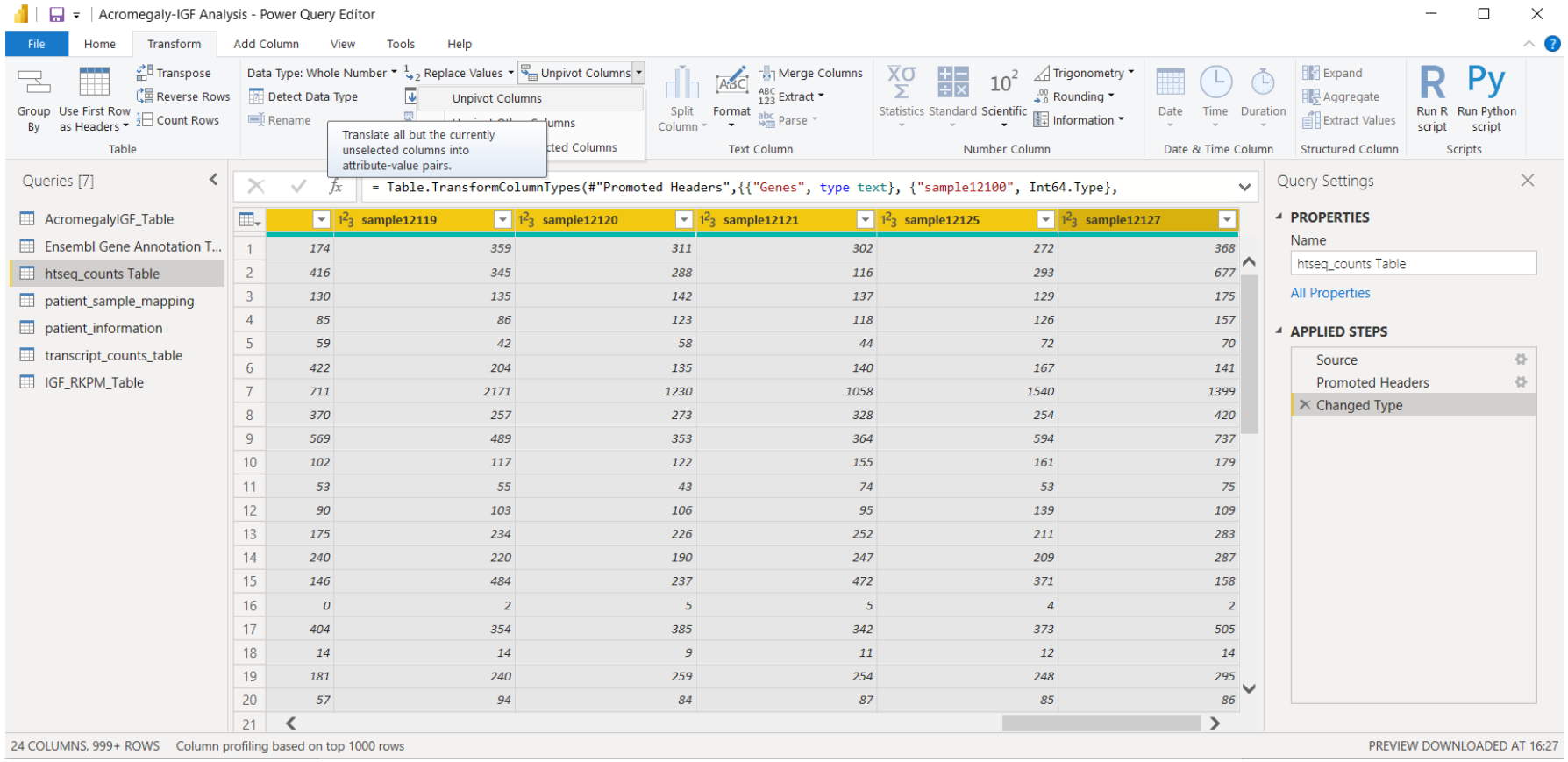
1.2. Edit Relationships

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
OK
Cancel

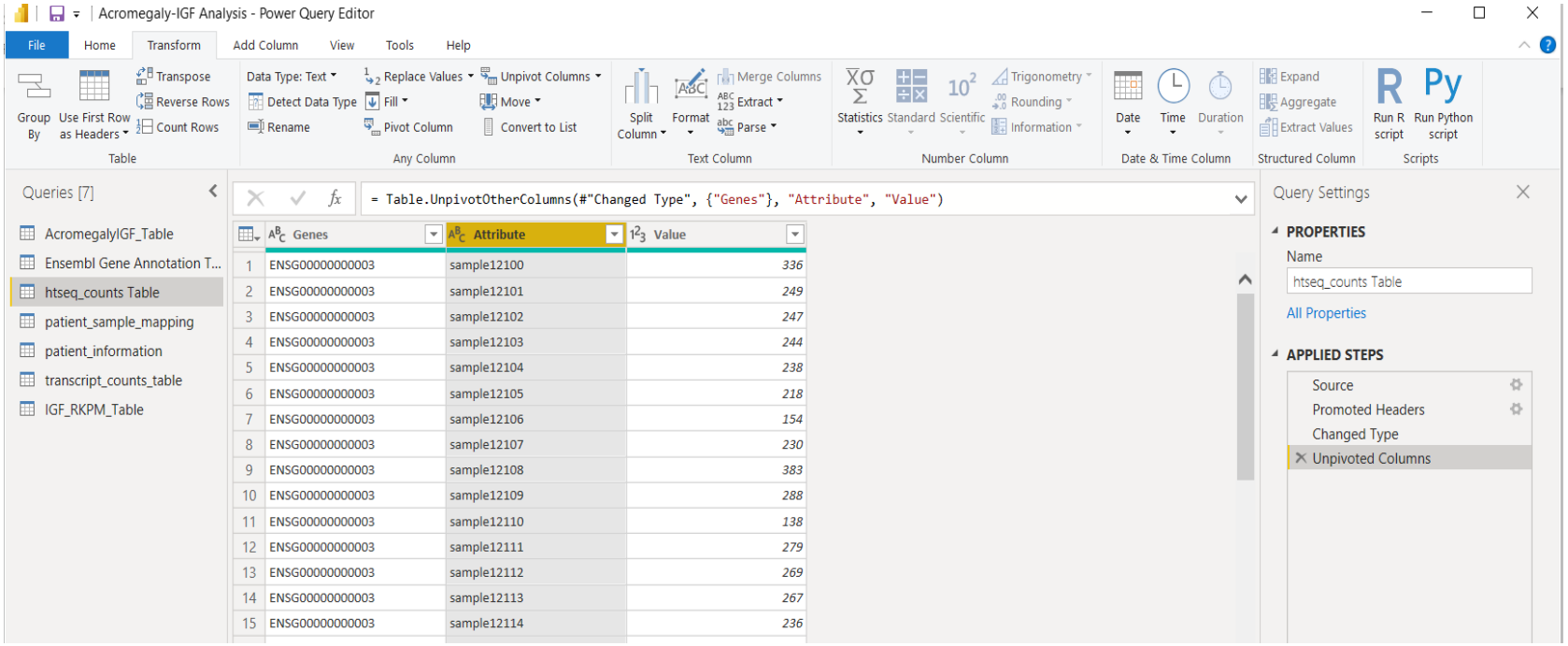
1.3. Unpivoting Columns

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



	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						

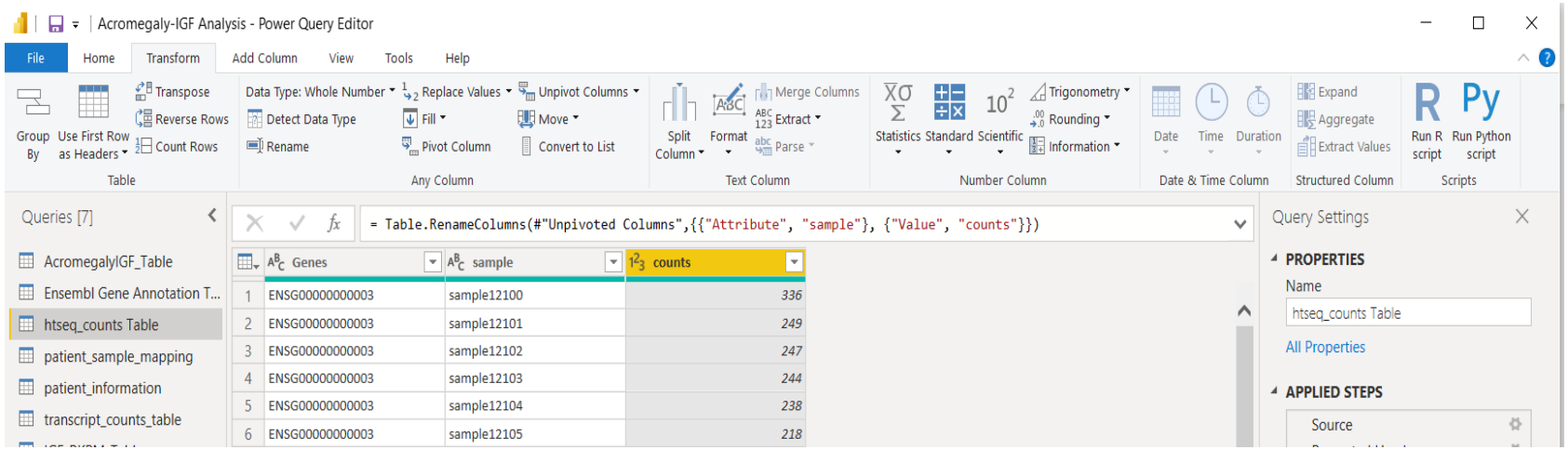
Table transforms as below, now rename the **Attribute** and **Value** Columns to **sample** and **counts** respectively.



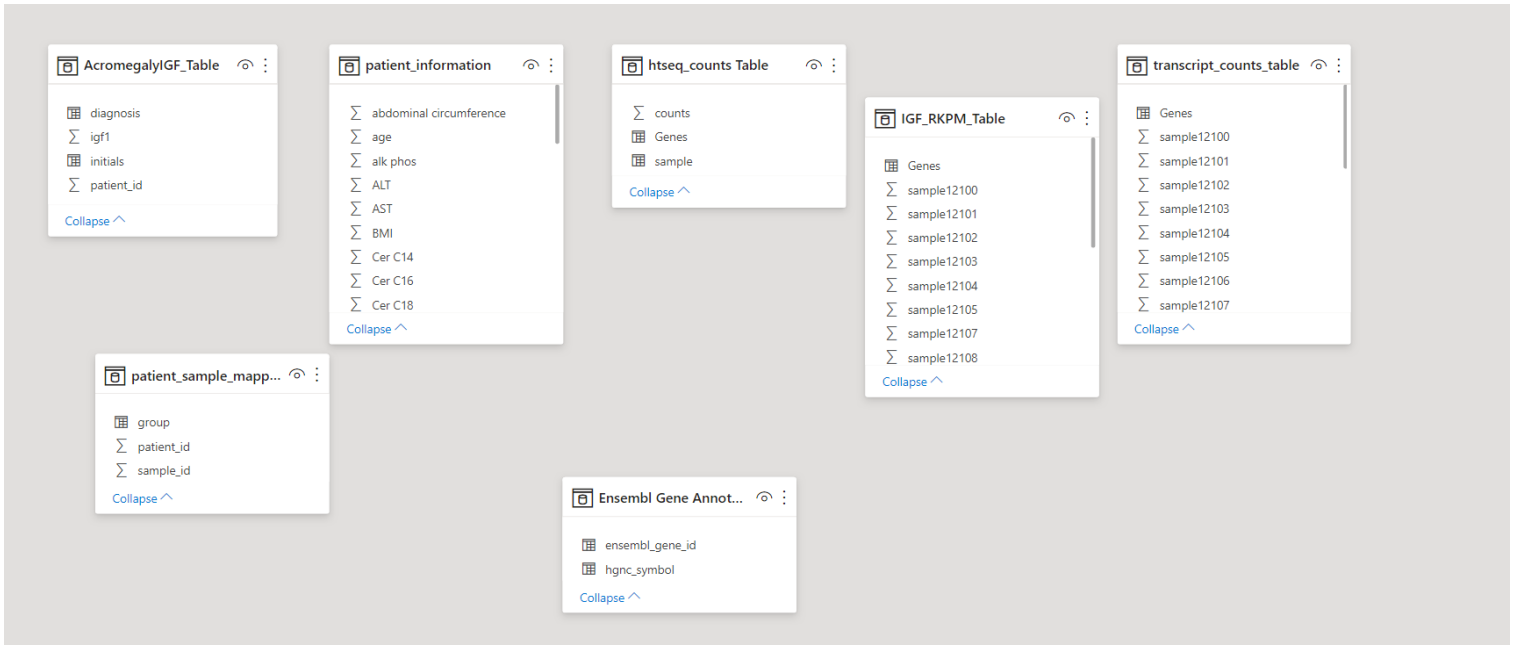
1.4. Renaming Columns

Renaming Attribute and Value Column using M formula,

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



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

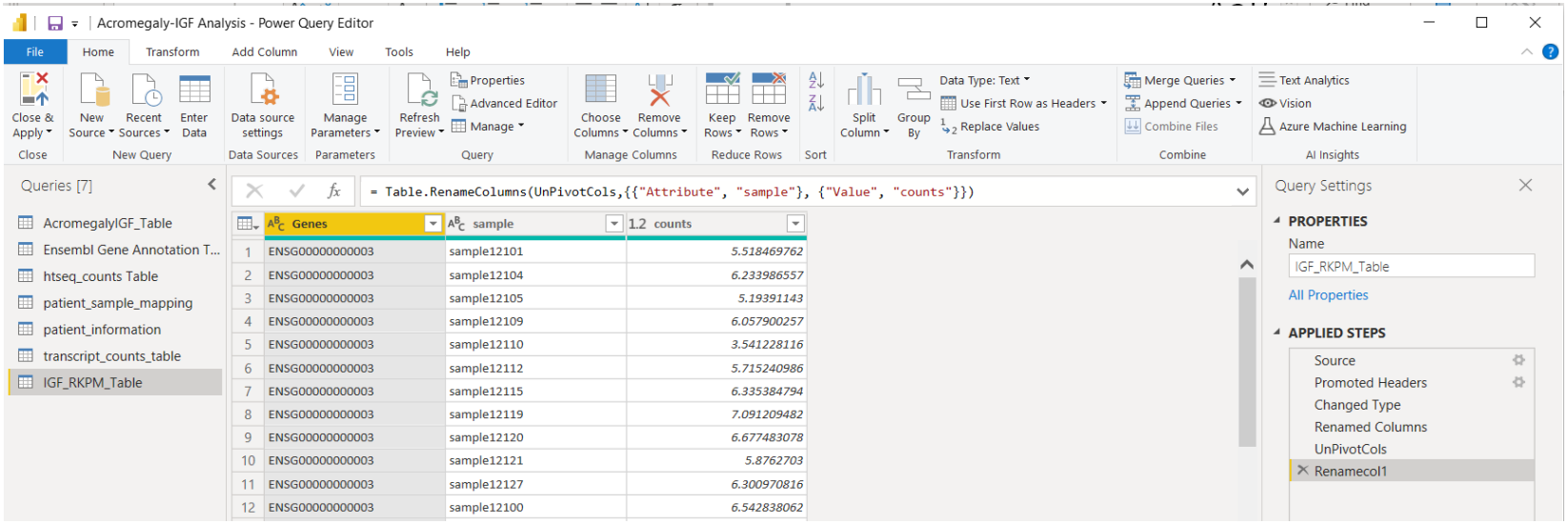


Repeating the unpivot and rename column steps in **IGF_RKPM_Table**.

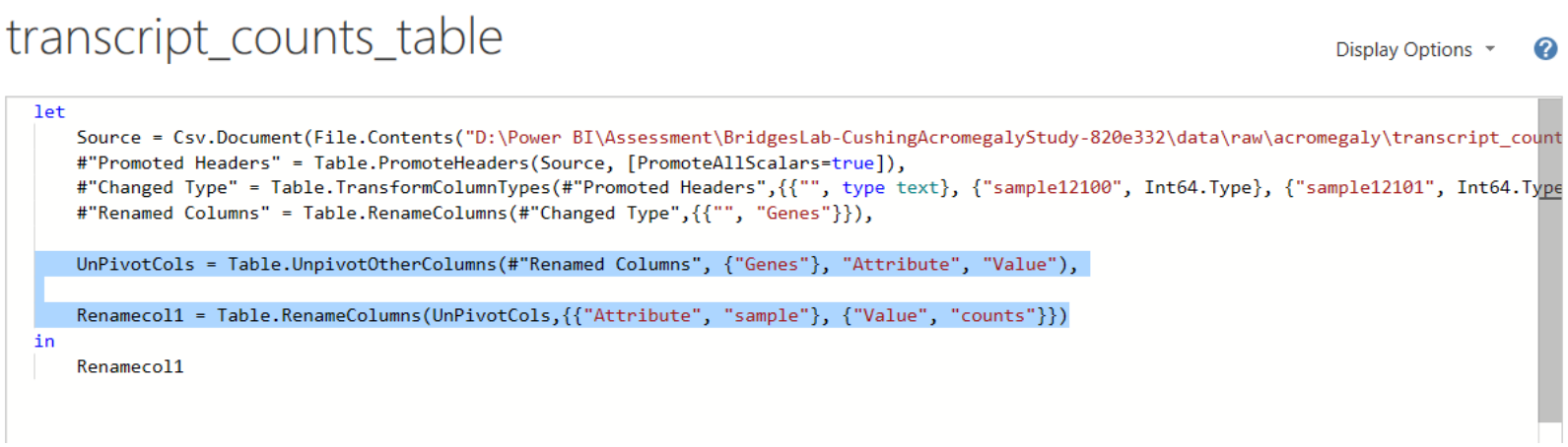
IGF_RKPM_Table

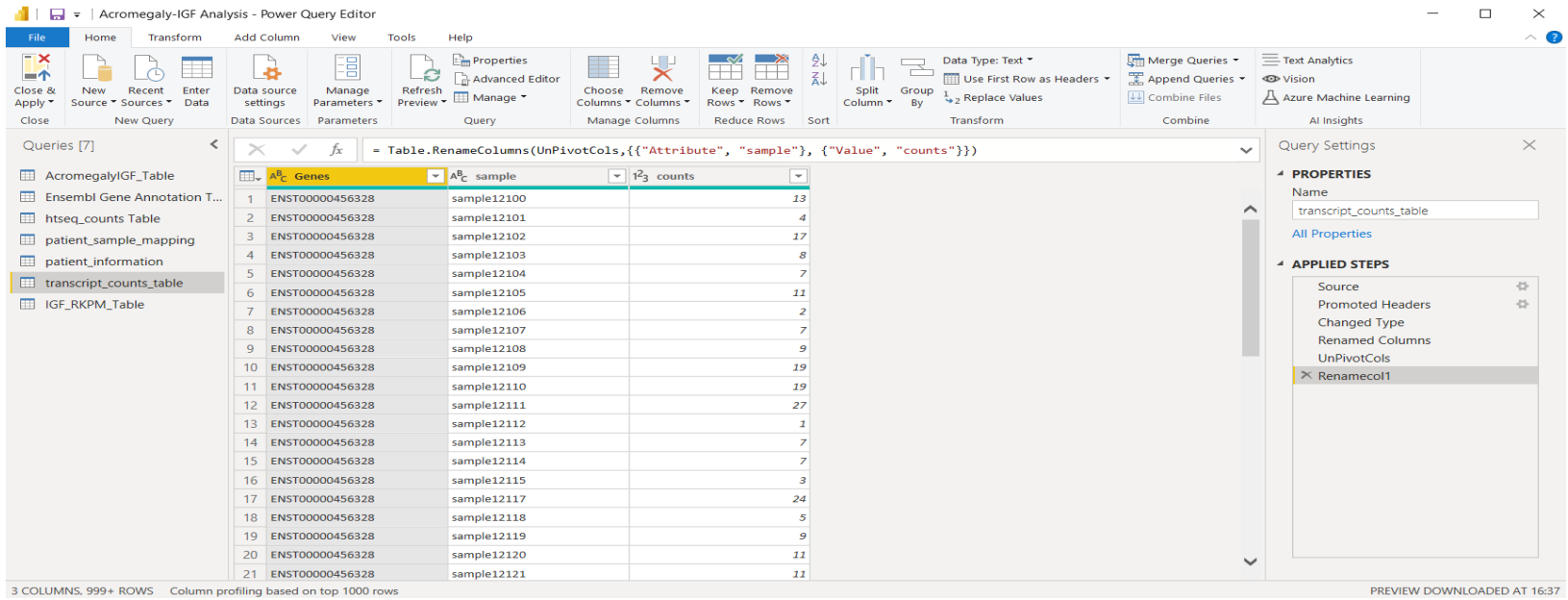
Display Options

```
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
```

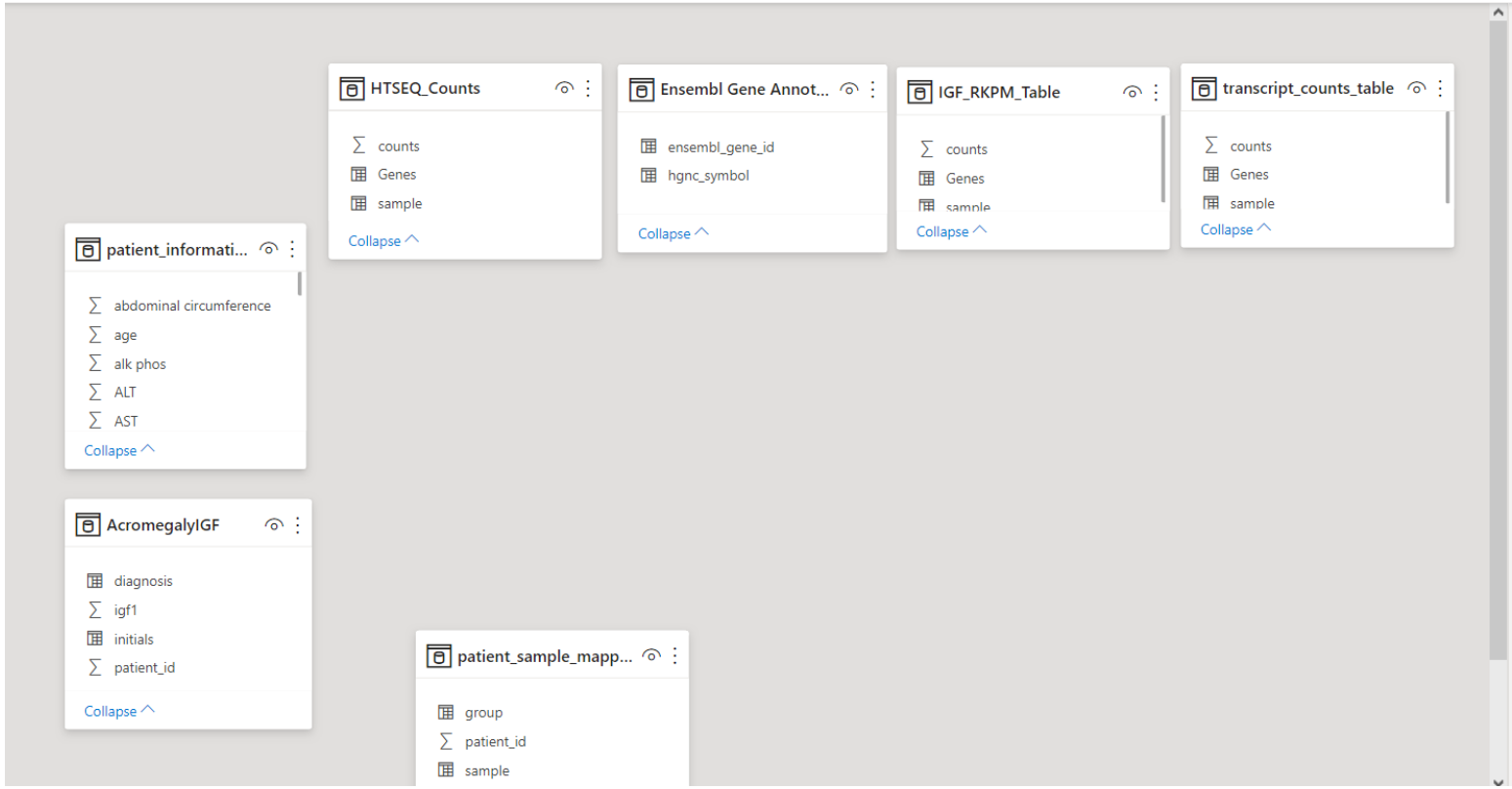


Repeating the unpivot and rename column steps in **transcript_counts_Table**.



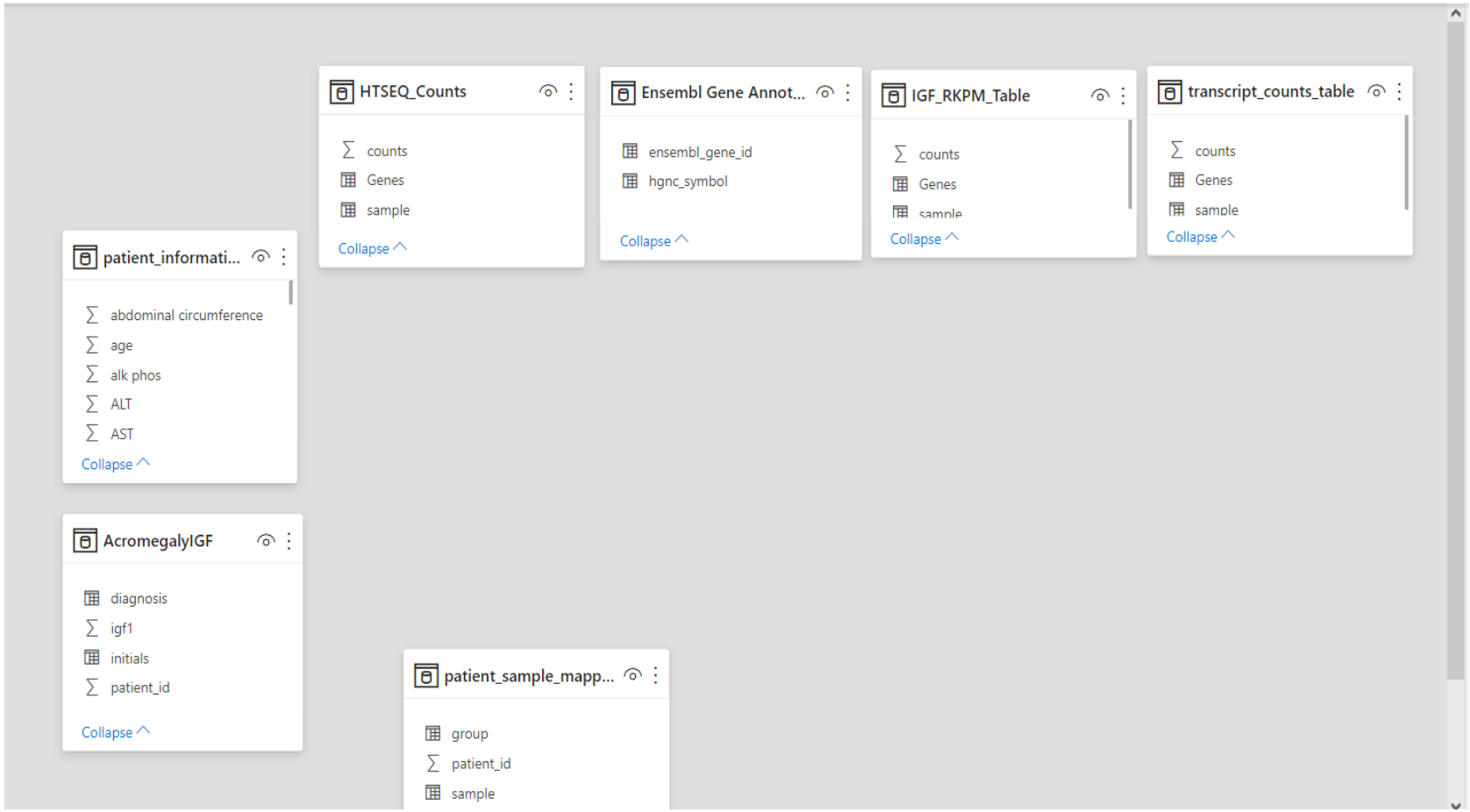


The model view of the data table is as below.


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

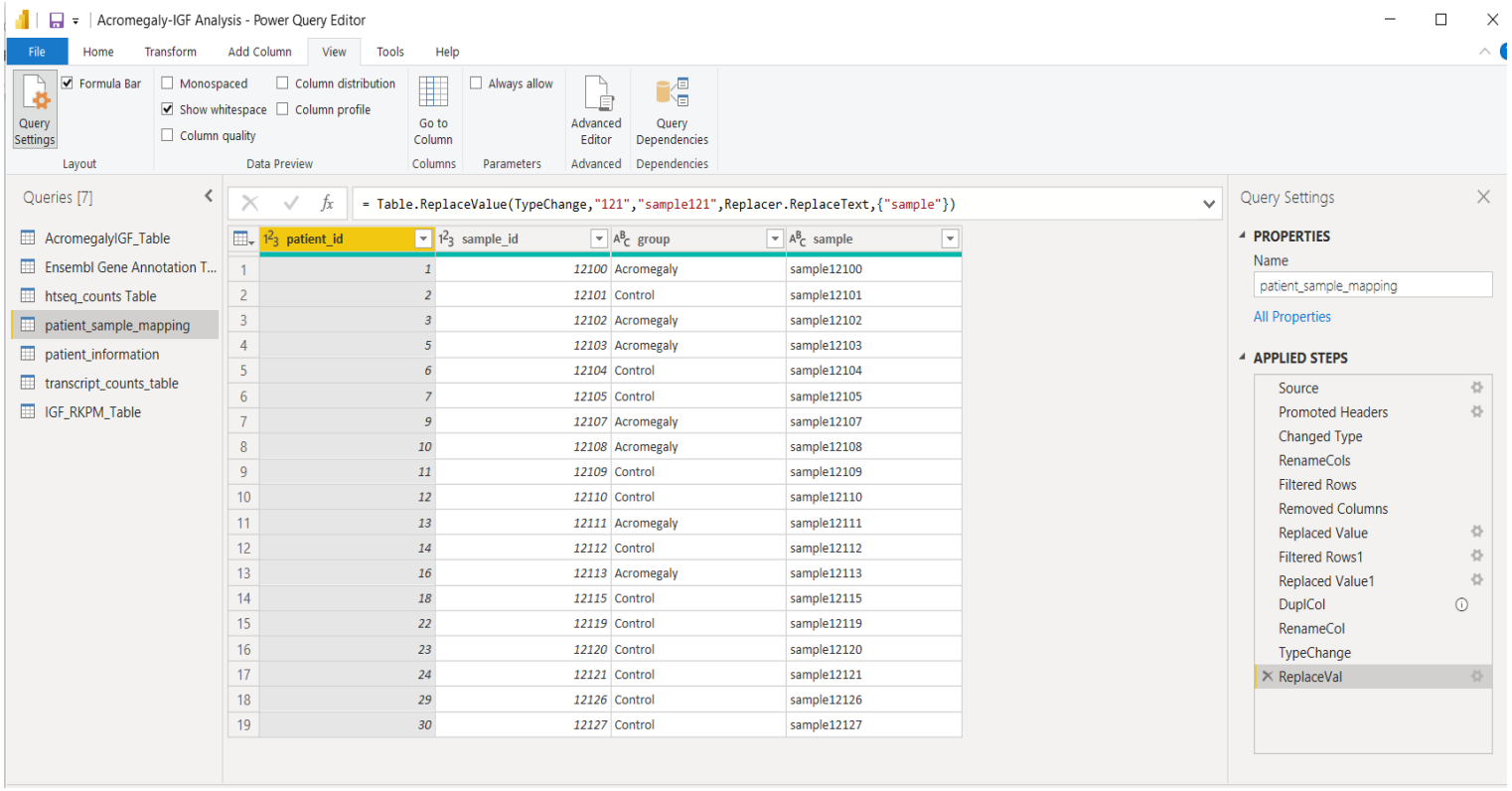
2. Star Schema / Snowflake Schema – Facts and Dimensions

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

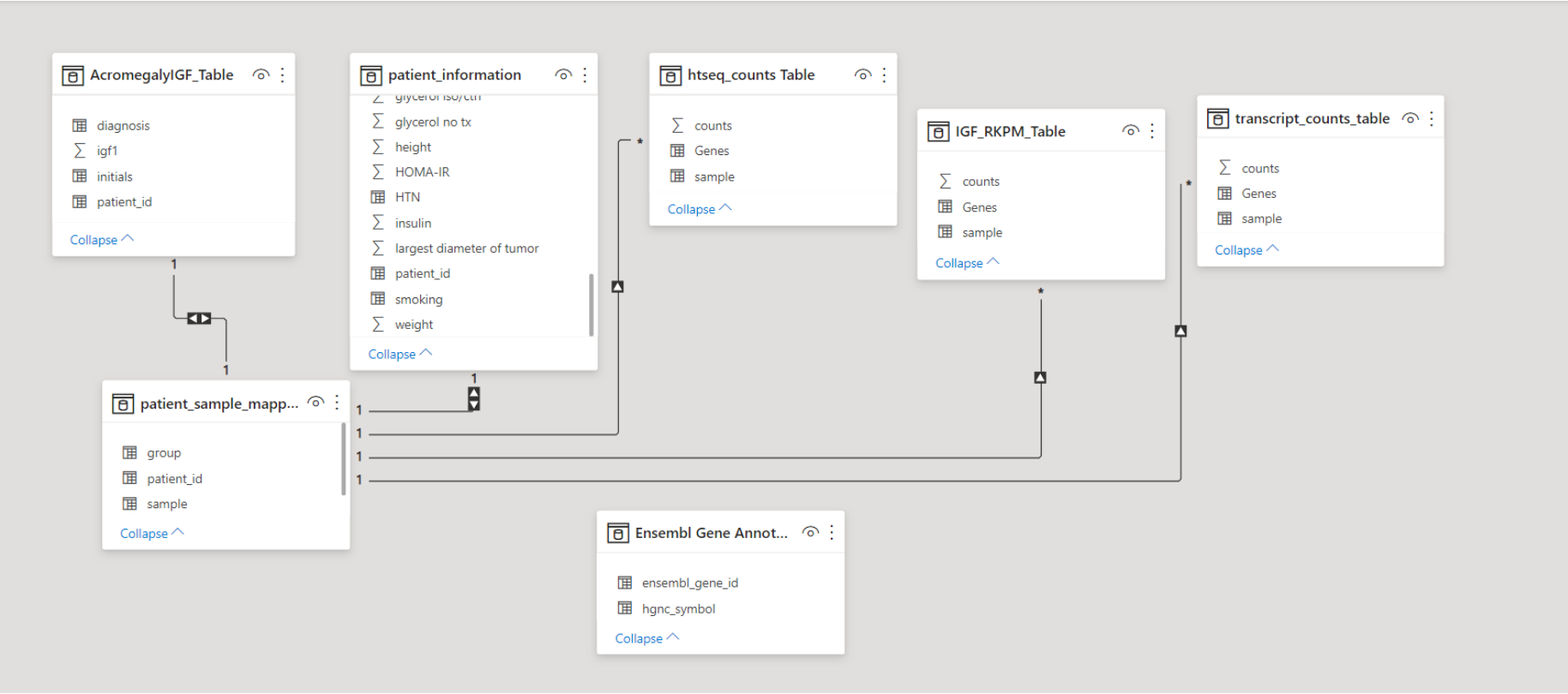


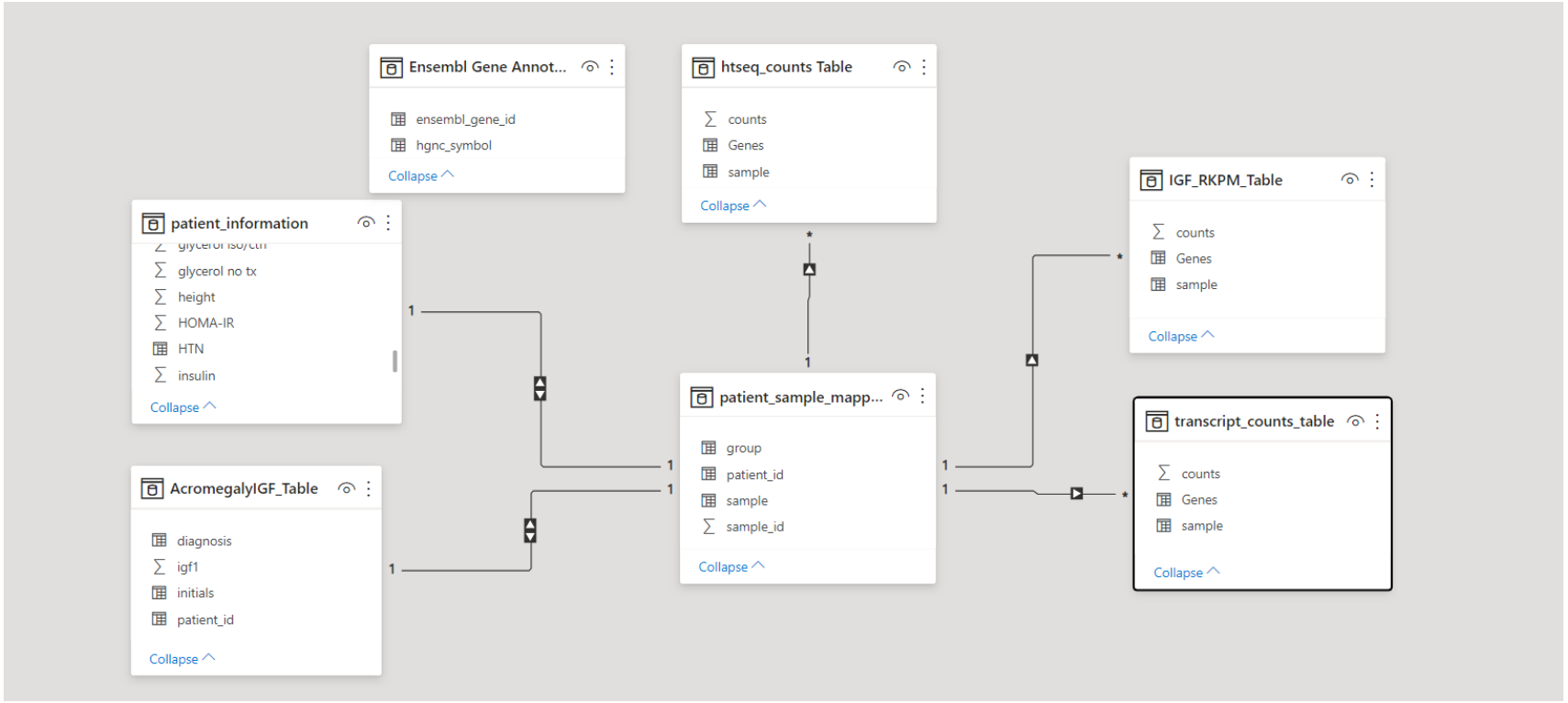
From the model view, it is evident that **the patient_sample_maping** table is the connecting table between all the other tables. Patient information and acromegaly patient-related data are stored in the tables with **patient_id** as the key. The gene_id, sequence counts in the other tables are related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


Acromegaly-IGF Analysis - Power Query Editor
File
Home
Transform
Add Column
View
Tools
Help
Formula Bar
Monospaced
Column distribution
Show whitespace
Column profile
Column quality
Query Settings
Layout
Data Preview
Go to Column
Columns
Always allow
Parameters
Advanced Editor
Advanced
Query Dependencies
Dependencies
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

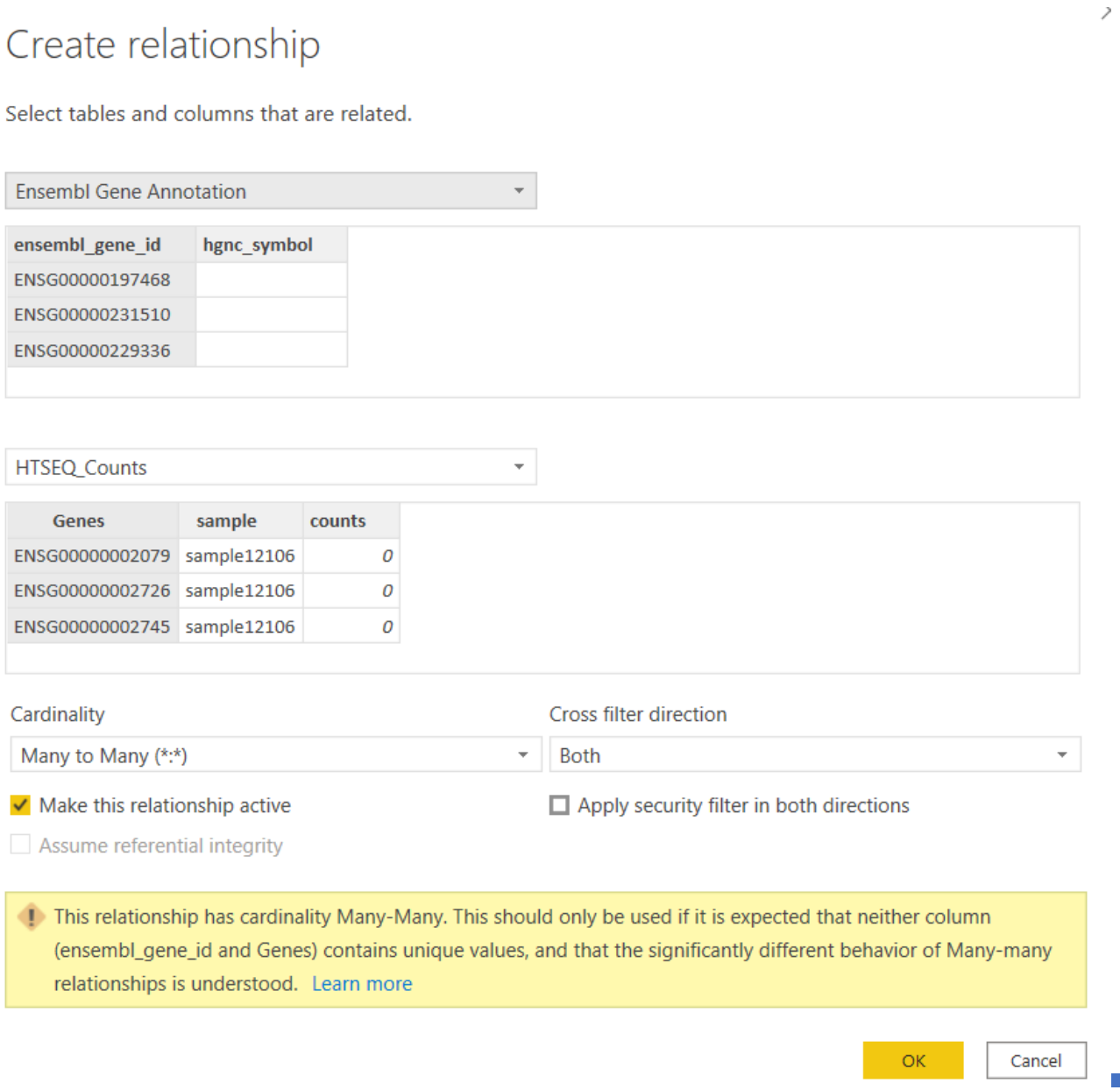
2.2. Fact Table

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.

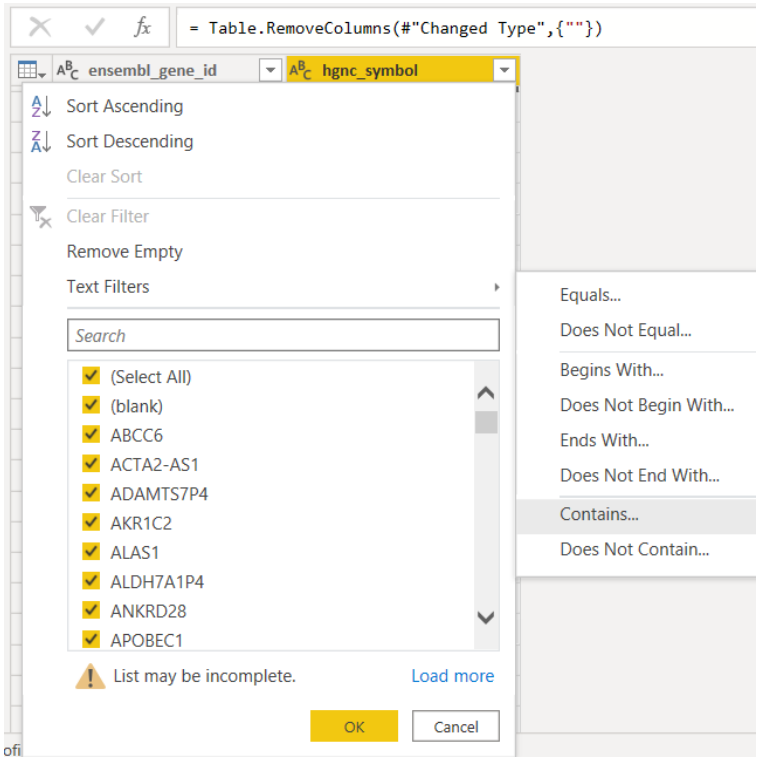

Create relationship
Select tables and columns that are related.
Ensembl Gene Annotation
ensembl_gene_id
hgnc_symbol
ENSG00000197468
ENSG00000231510
ENSG00000229336
HTSEQ_Counts
Genes
sample
counts
ENSG00000002079
sample12106
0
ENSG00000002726
sample12106
0
ENSG00000002745
sample12106
0
Cardinality
Many 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
OK
Cancel

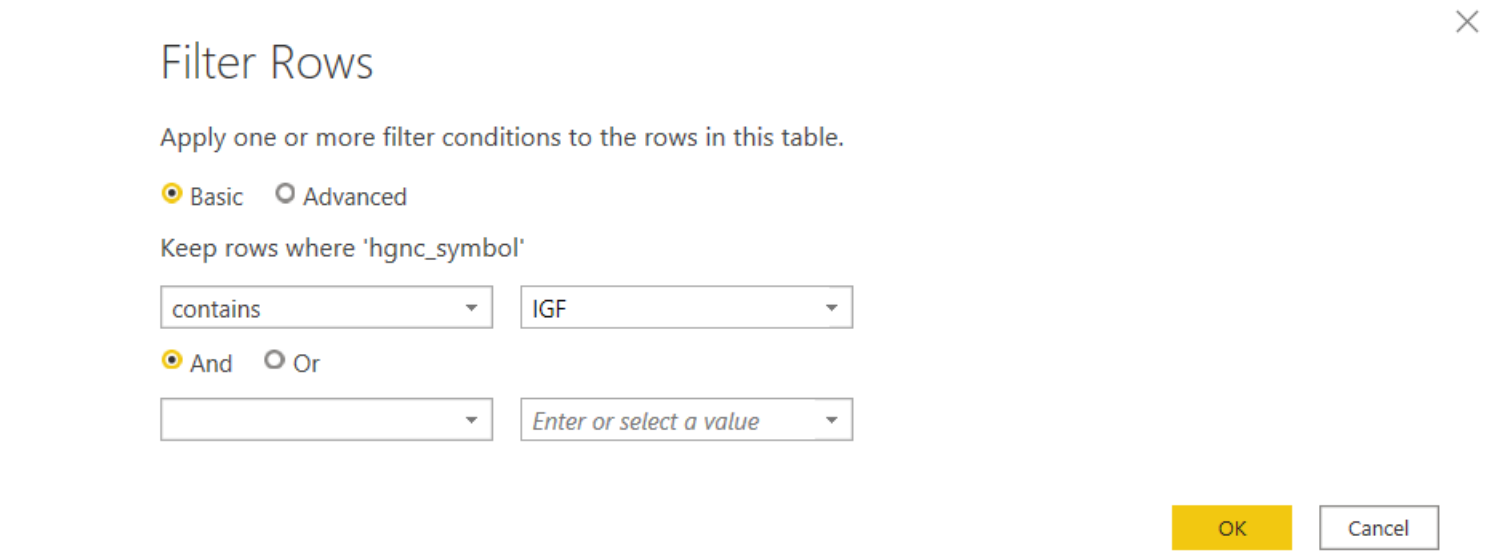
To avoid this many to many and connect the tables we need to follow 3 steps as below.

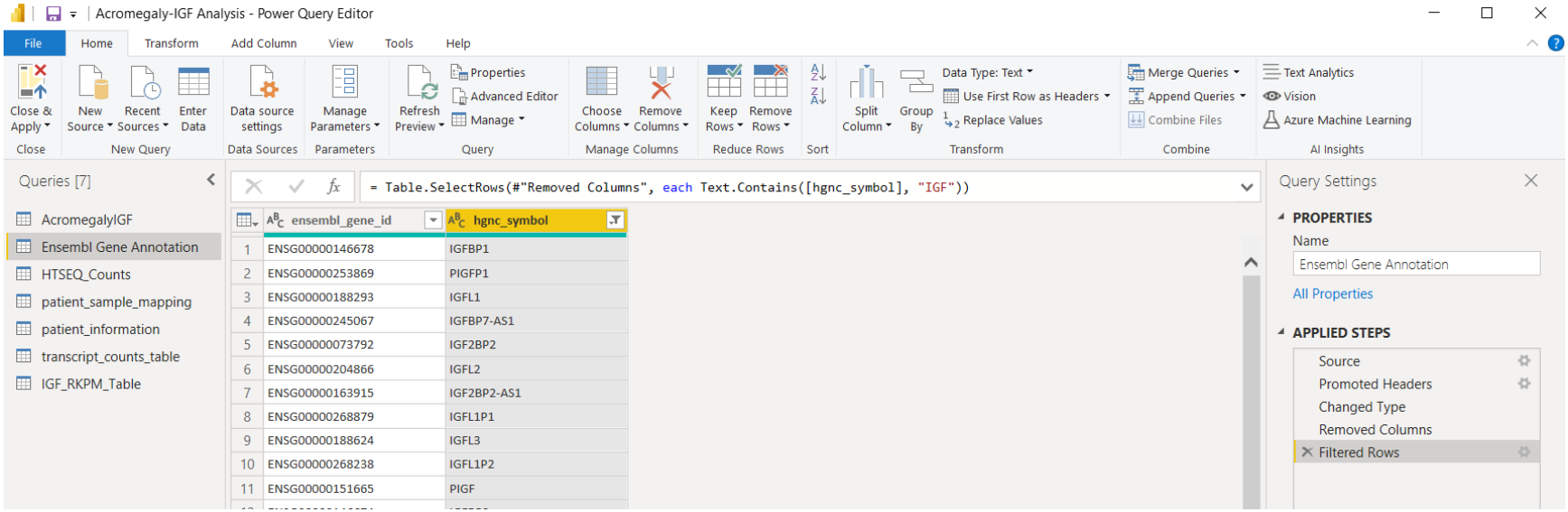
1. Filter Rows: We need only IGF gene-related Data, so we filter the Ensembl Gene Annotation table with hgnc_symbol containing “igf”
2. Merge Queries: merge htseq_counts Table with patient_sample_mapping and create a new column with diagnosis “group” details.
3. Create a New Relationship :

2.4. Filter Rows

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

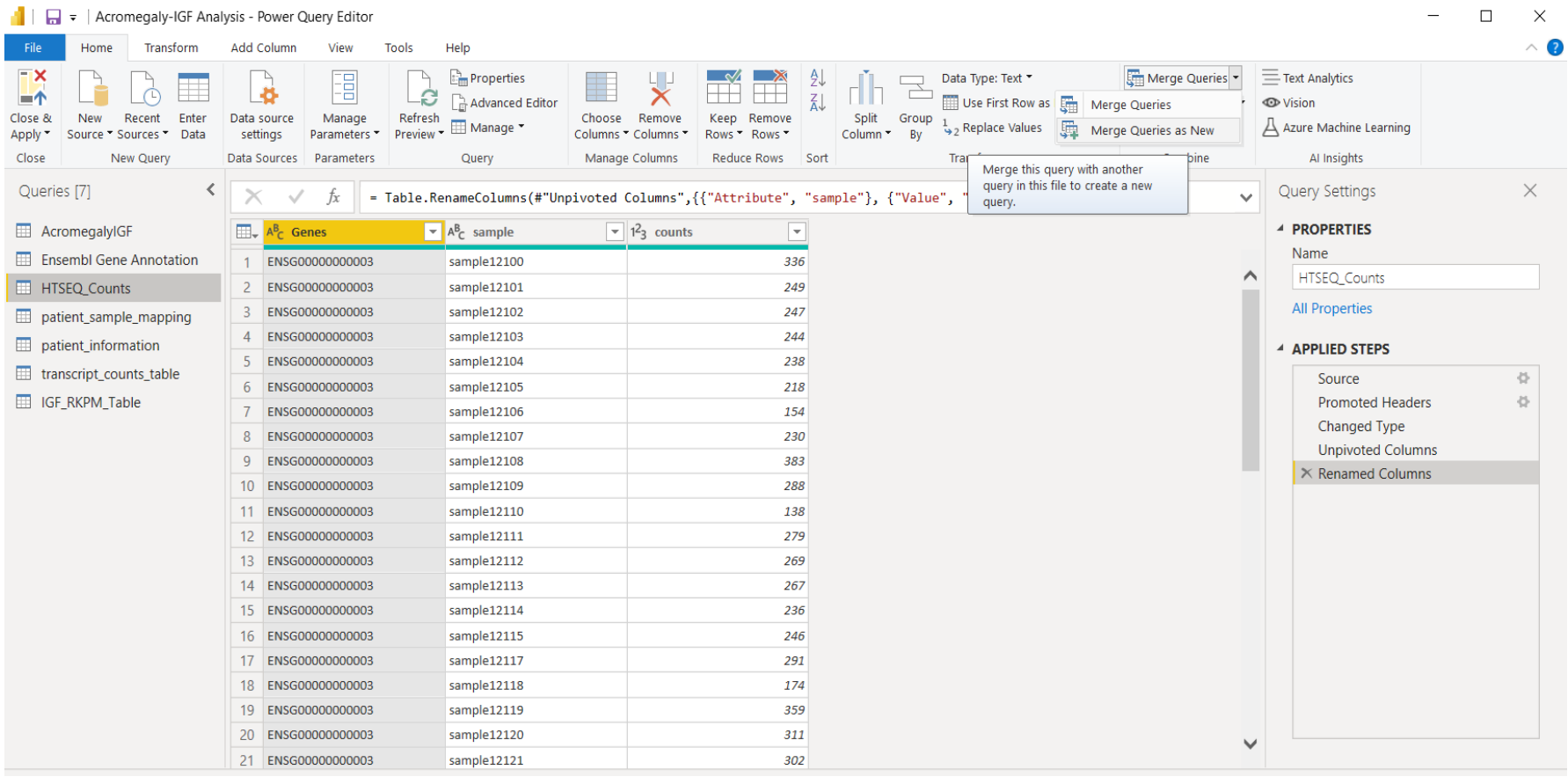


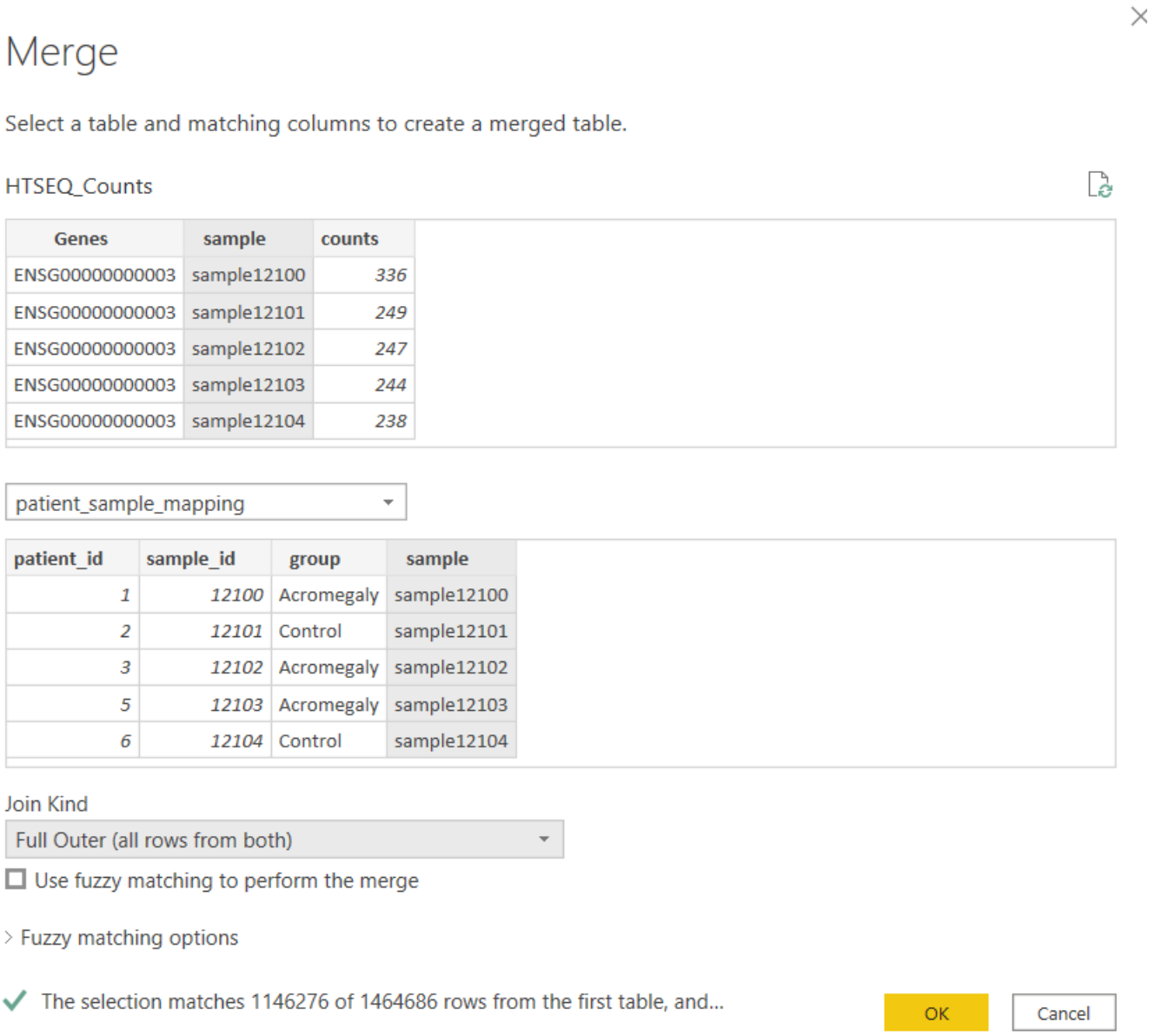

Filter Rows
Apply one or more filter conditions to the rows in this table.
Basic
Advanced
Keep rows where 'hgnc_symbol'
contains
IGF
And
Or
Enter or select a value
OK
Cancel


Acromegaly-IGF Analysis - Power Query Editor
File
Home
Transform
Add Column
View
Tools
Help
Close & Apply
New Source
Recent Sources
Enter Data
Data source settings
Manage Parameters
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Merge Queries
Append Queries
Combine Files
Text Analytics
Vision
Azure Machine Learning
Queries [7]
AcromegalyIGF
Ensembl Gene Annotation
HTSEQ_Counts
patient_sample_mapping
patient_information
transcript_counts_table
IGF_RKPM_Table
= 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

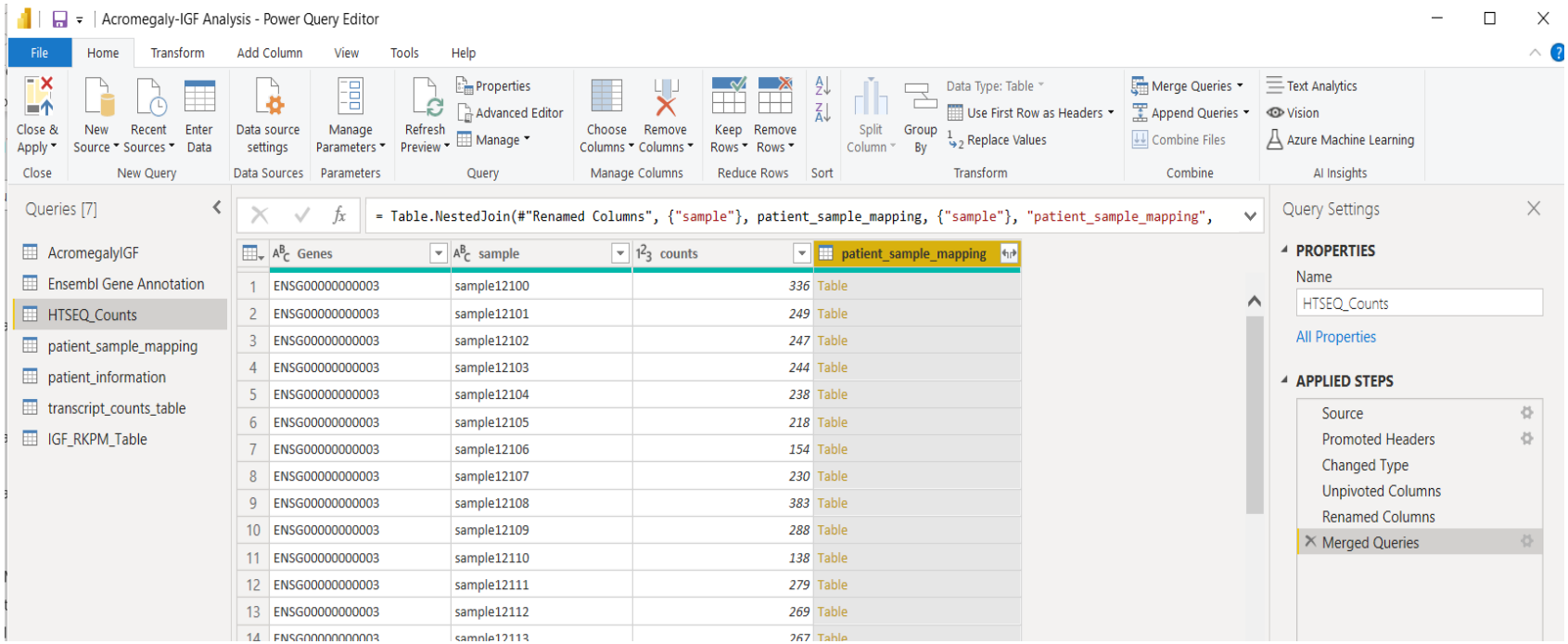
Merging htseq_counts Table with patient_sample_mapping and create a new column with diagnosis “group” details.

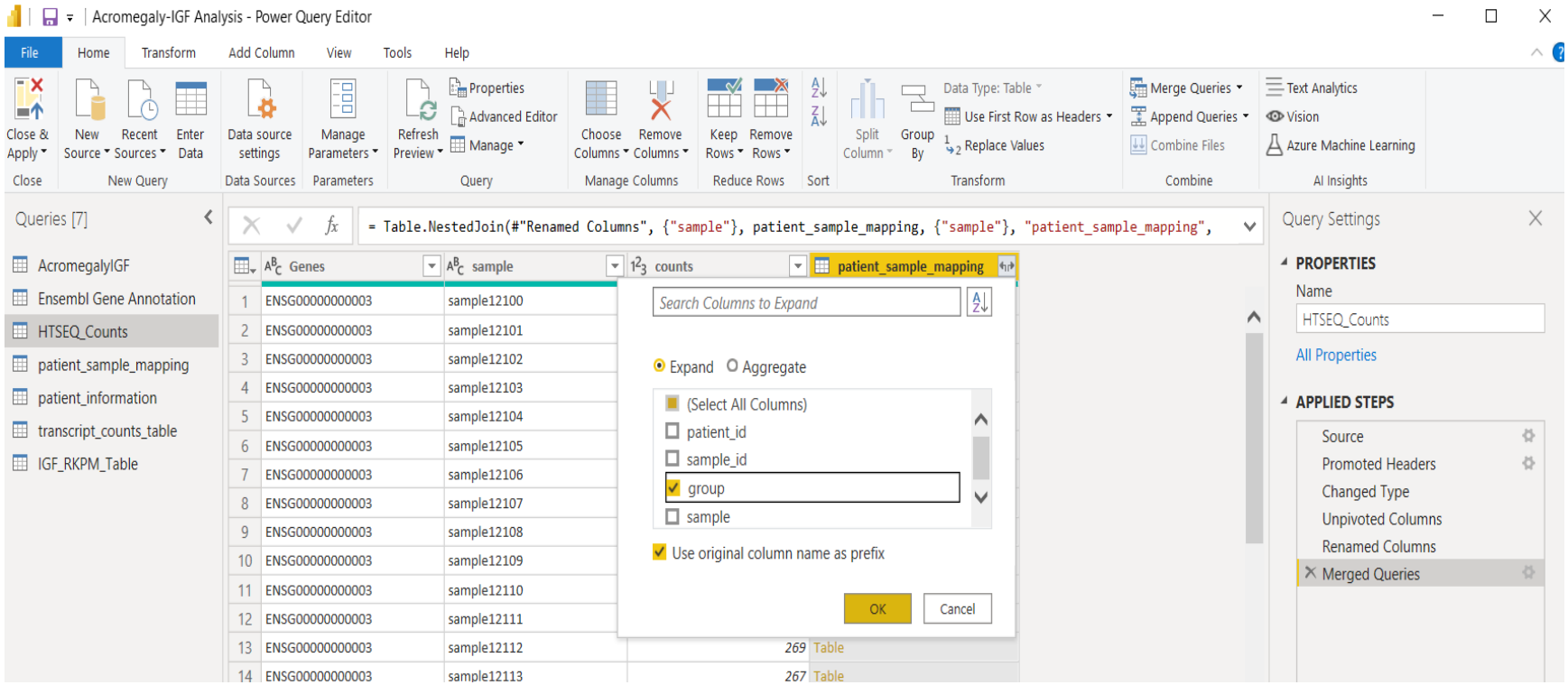


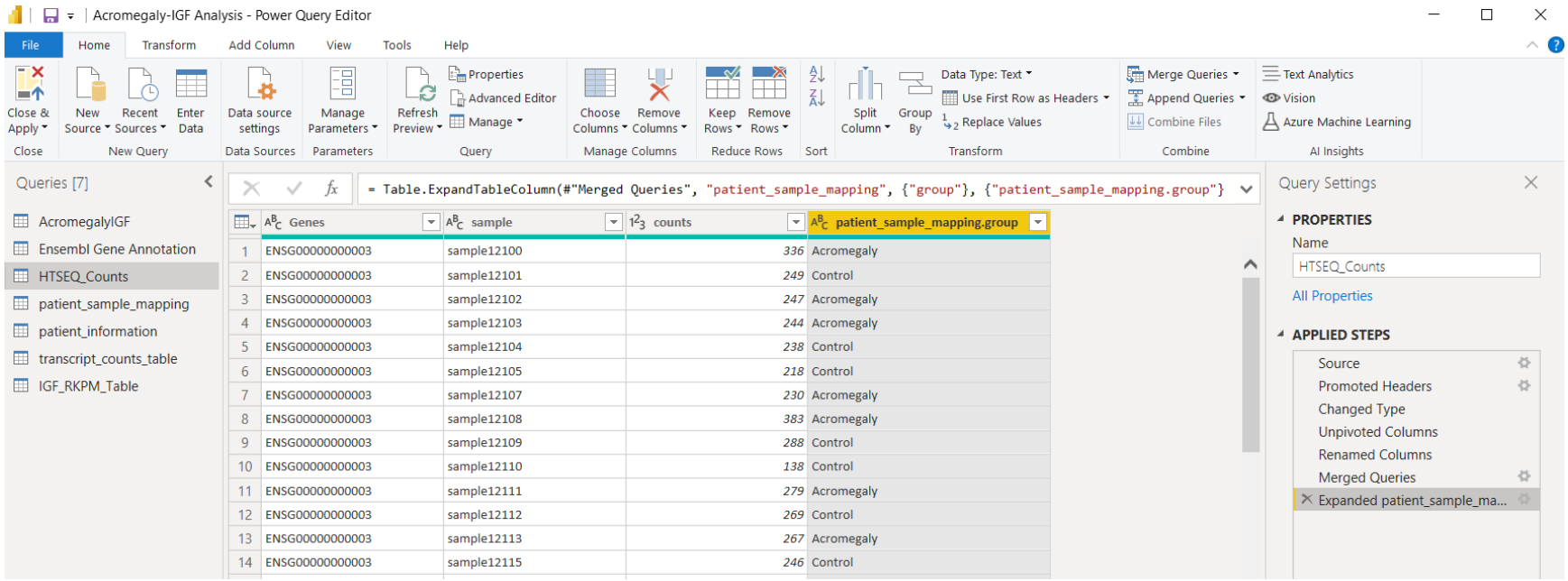


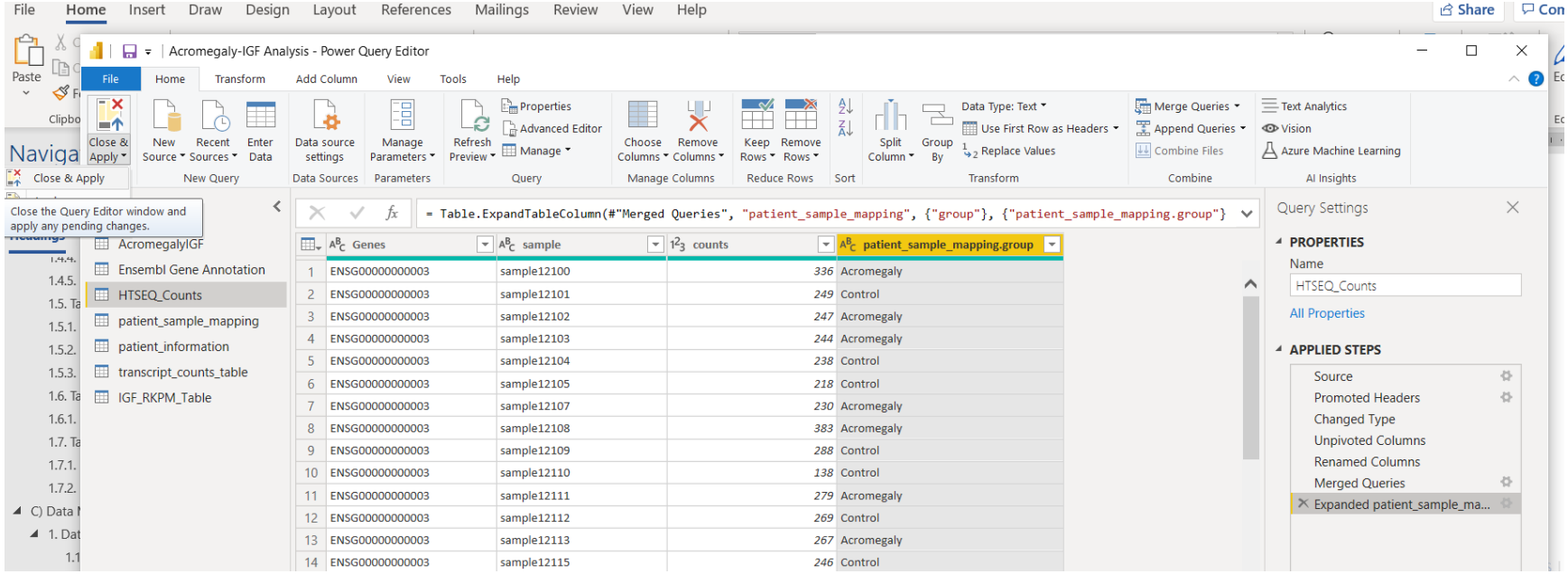
Selecting group Column from

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



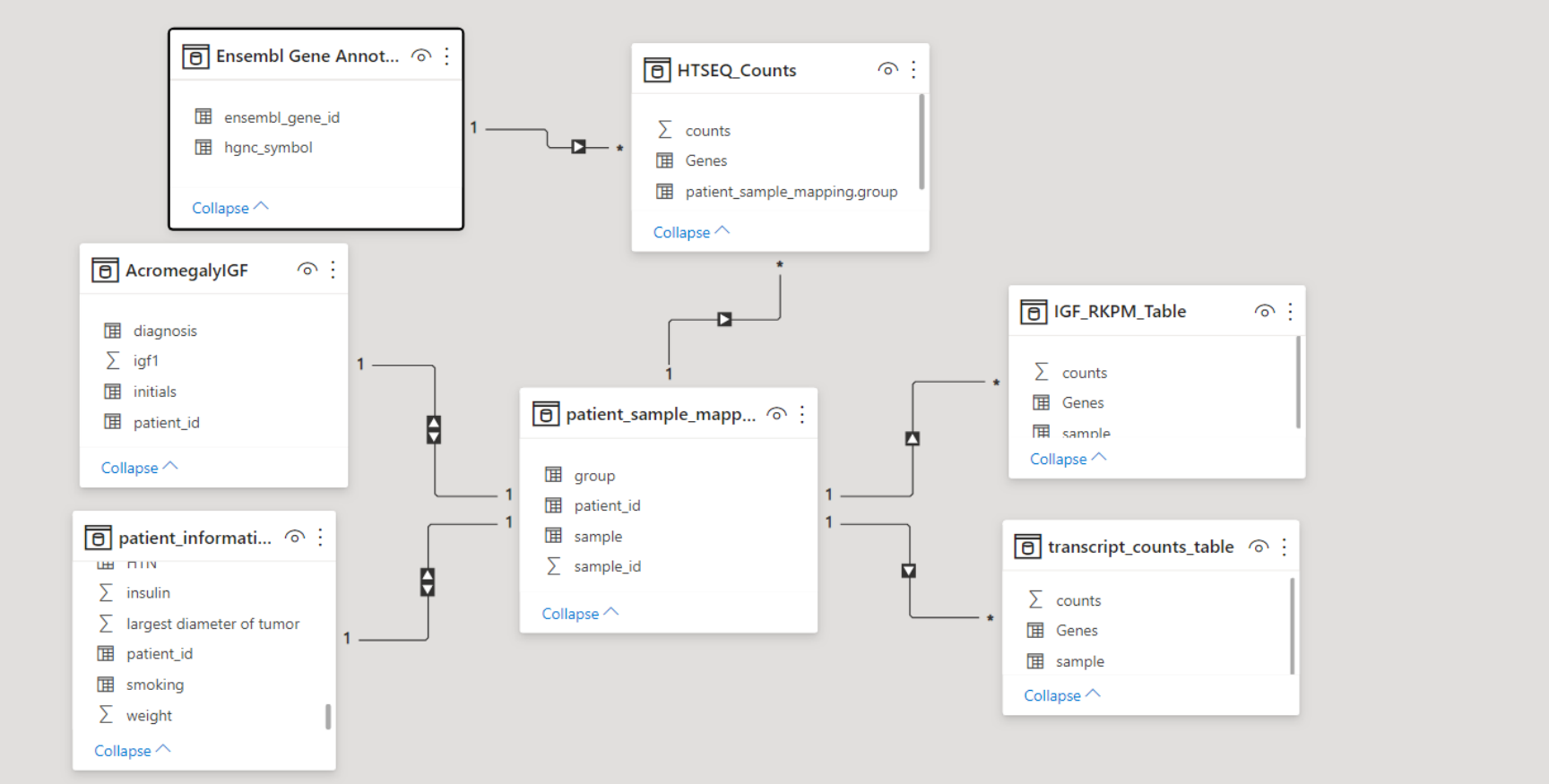

Acromegaly-IGF Analysis - Power Query Editor
File
Home
Transform
Add Column
View
Tools
Help
Close & Apply
New Source
Recent Sources
Enter Data
Data source settings
Manage Parameters
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Split Column
Group By
Data Type: Table
Use First Row as Headers
Replace Values
Merge Queries
Append Queries
Combine Files
Text Analytics
Vision
Azure Machine Learning
Close
New Query
Data Sources
Parameters
Query
Manage Columns
Reduce Rows
Sort
Transform
Combine
AI Insights
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
ENSG00000000003
sample12100
sample12101
sample12102
sample12103
sample12104
sample12105
sample12106
sample12107
sample12108
sample12109
sample12110
sample12111
sample12112
sample12113
269
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


Acromegaly-IGF Analysis - Power Query Editor
File
Home
Transform
Add Column
View
Tools
Help
Close & Apply
New Source
Recent Sources
Enter Data
Data source settings
Manage Parameters
Refresh Preview
Properties
Advanced Editor
Manage
Choose Columns
Remove Columns
Keep Rows
Remove Rows
Split Column
Group By
Data Type: Text
Use First Row as Headers
Replace Values
Merge Queries
Append Queries
Combine Files
Text Analytics
Vision
Azure Machine Learning
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
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...

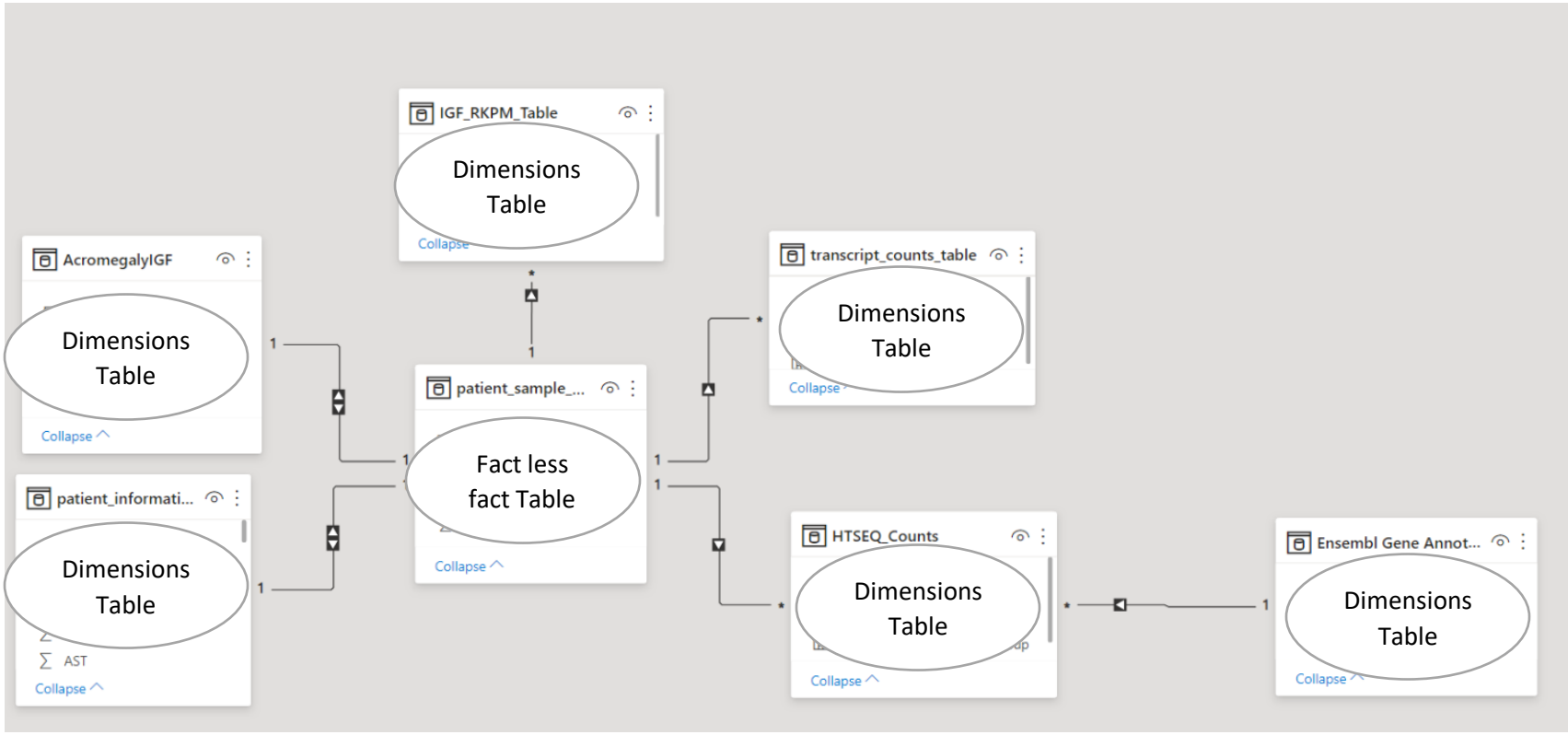

Acromegaly-IGF Analysis - Power Query Editor
Close & Apply
Close the Query Editor window and apply any pending changes.
AcromegalyIGF
Ensembl Gene Annotation
HTSEQ_Counts
patient_sample_mapping
patient_information
transcript_counts_table
IGF_RKPM_Table
= Table.ExpandTableColumn(#"Merged Queries", "patient_sample_mapping", {"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...

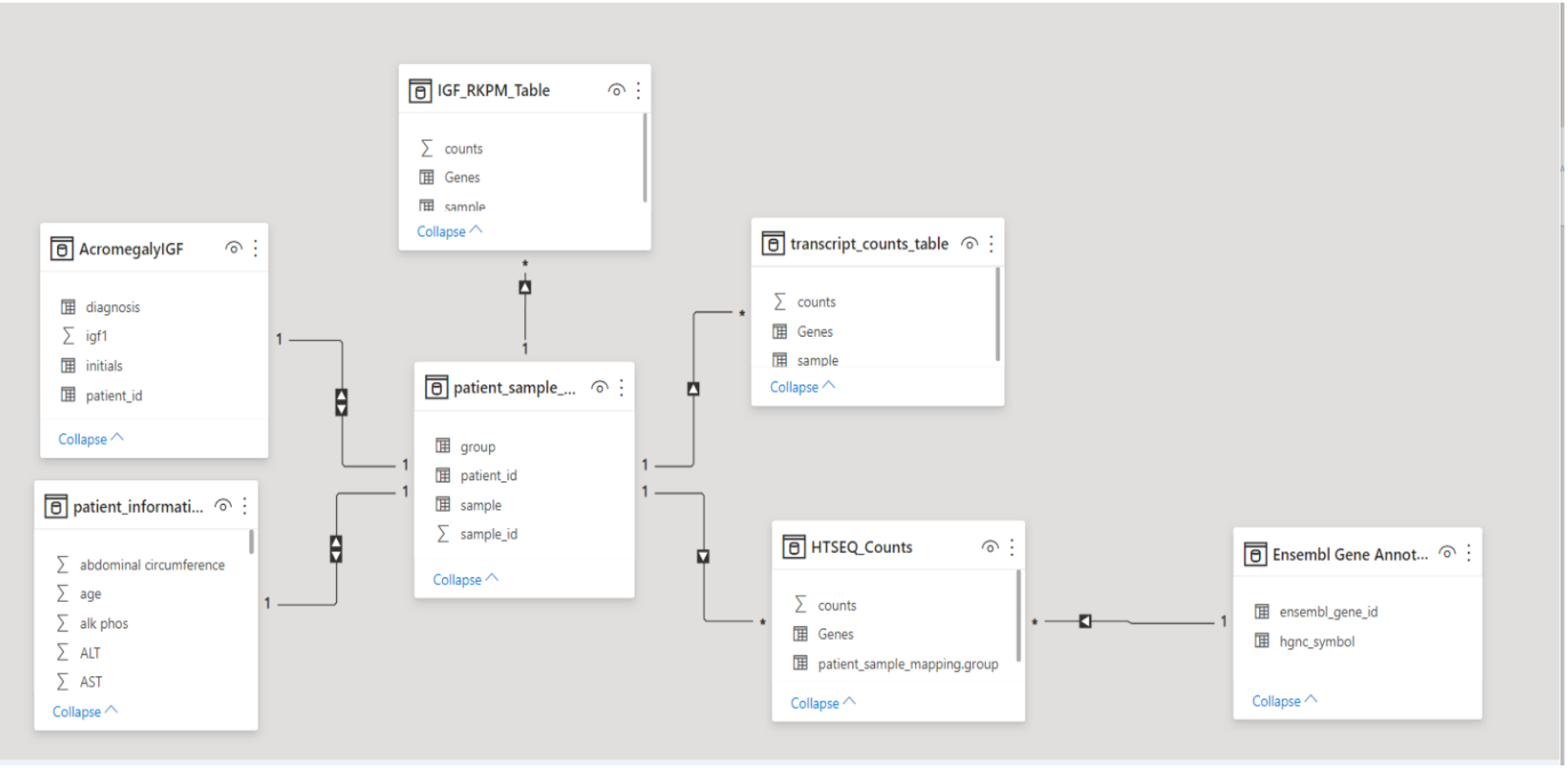
2.6. Create a New Relationship

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
AST
Collapse
HTSEQ_Counts
Dimensions Table
Collapse
Ensembl Gene Annot...
Dimensions Table
Collapse
1
*


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

BIG Data and Business Intelligence
In-Course Assessment

ACROMEGALY AND IGF ANALYSIS

Section 2: Business Intelligence Solution

15/01/2021

STUDENT

Mohana Kamanooru

A0223038@tees.ac.uk

MODULE LEADER

Dr. Annalisa Occhipinti

a.occhipinti@tees.ac.uk

Executive Summary

Abstract

Acromegaly is a pituitary tumor that produces high levels of growth hormone. This research is to understand and analyze how high the IGF levels could rise. If there is any resistance to insulin and blood glucose levels and does the physical appearance and physical characteristics of patients change when affected by Acromegaly? The data set is acquired from opensource. The dataset is evaluated and preprocessed to build the appropriate data model. The transformed data model is then used to draw solutions for the mentioned business questions. From this analysis, it can be stated that Acromegaly causes a rise in IGF levels, higher insulin resistance, higher glucocorticoid levels, and does not have significant changes concerning BMI, Age, and height. Except there was a marginal increase observed in height and BMI.

Data Model

This secondary research dataset contains the details gathered from 9 Acromegaly patients and 11 control patients. The gene sequences are generated and analyzed. The data model is a snowflake schema with a factless fact table as shown in the picture below.

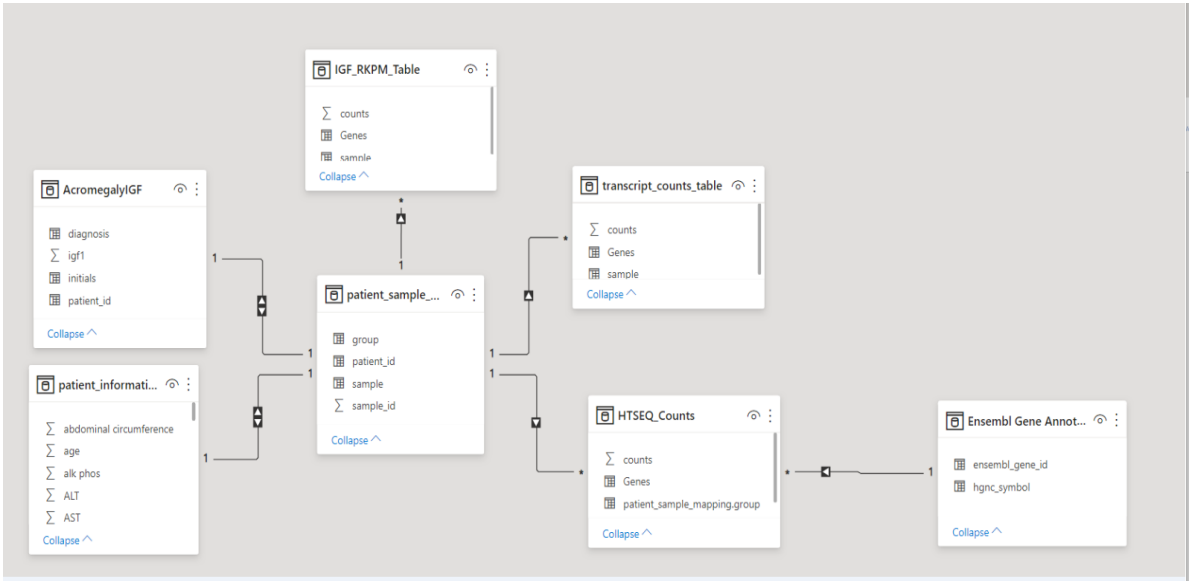
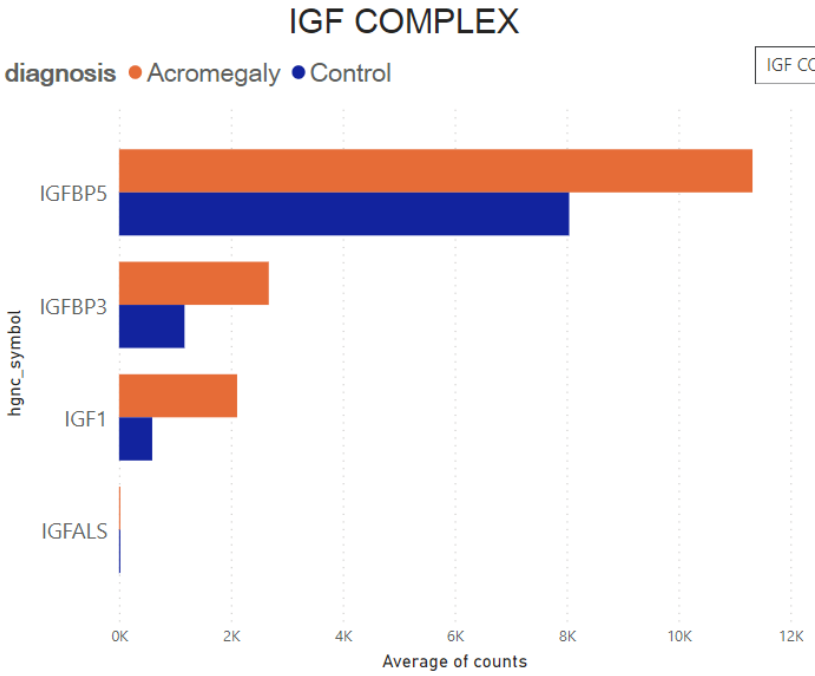
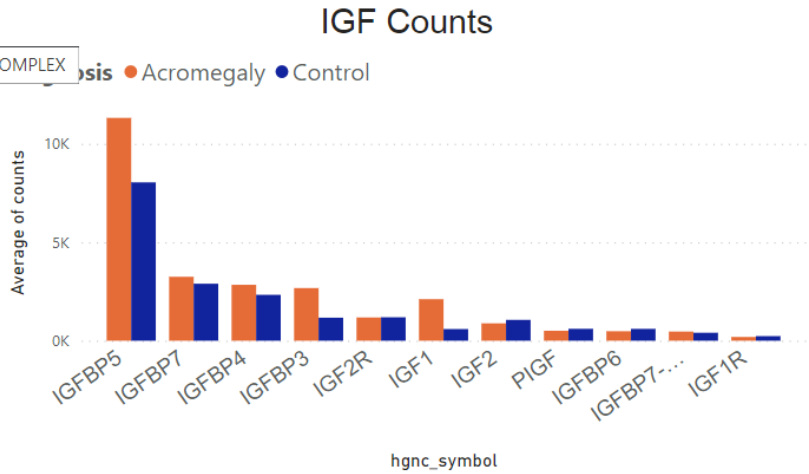


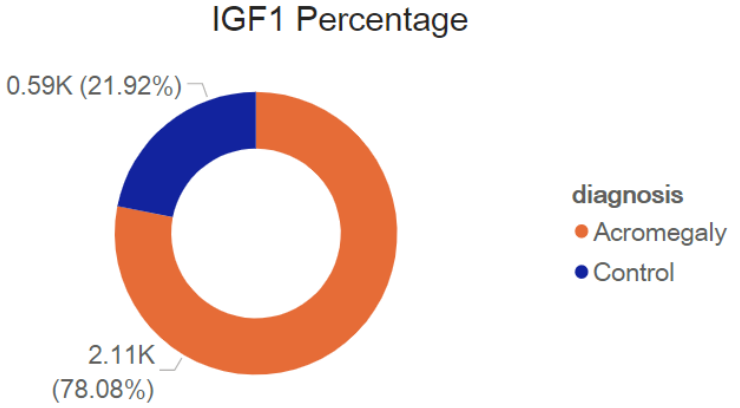
Figure 1 Data Model - Snowflake Schema

Key Findings

Growth hormone stimulates cell reproduction, cell regeneration, and is also responsible for growth in the body. In deeper terms, the pituitary gland releases IGF1, and growth is the result of circulating this IGF1 in the body. Below are the plots between IGF components and the sequence counts.







2,110.00	592.36	1518
Acromegaly IGF1 Levels	Control IGF1 Levels	Difference in IGF1 Levels
2,675.14	1,169.73	1,505.42
Acromegaly IGFBP3 Levels	Control IGFBP3 Levels	Difference in IGFBP3 Levels

From these plots, it is evident that the IGF levels have a rapid increase in the IGF levels. IGF complex functions show an enormous increase in IGFBP5, IGFBP3, and IGF. The donut plot shows the percentage increase, which is almost 3.5 times higher than the controls.

Conclusions

The following observations and conclusions have been made from the research.

1. The average IGF1 levels are 3.5 times higher in Acromegaly patients.
2. Acromegaly patients have higher insulin resistance and low levels of insulin sensitivity.
3. Acromegaly patients tend to show higher averages for BMI and height.
4. The probability of Acromegaly might increase with age. Since the analyzed data is a smaller sample, this may not be true for more significant models and real-time scenarios.
5. Weight and Abdominal circumferences do not tend to change in considerable amounts according to the research dataset.

Acknowledgments

Thanks to Professor Dr. Annalisa Occhipinti for her continual guidance, support, and advice throughout my research. I want to thank my parents Mr. Ramanachari Kamanooru, Mrs. Vijayalakshmi Kamanooru, and my 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.

Contents

Business Intelligence Solution

Introduction

Acromegaly is a rare pituitary tumour that may either cause gigantism or dwarfism. Acromegaly might also result in shortening the life expectancy of a person, other symptoms and side effects vary from short-term to long-term and are good in a number. Therefore, it is vital to understand the tumour and its symptoms. In this secondary research, the analyzed data is collected from different patients with and without pituitary tumour. The gene sequence counts, observations on patients are compared between the two categories to determine any relationship or change between age, BMI, IGF, insulin levels, and the adenoma.

Analysis and Evaluation

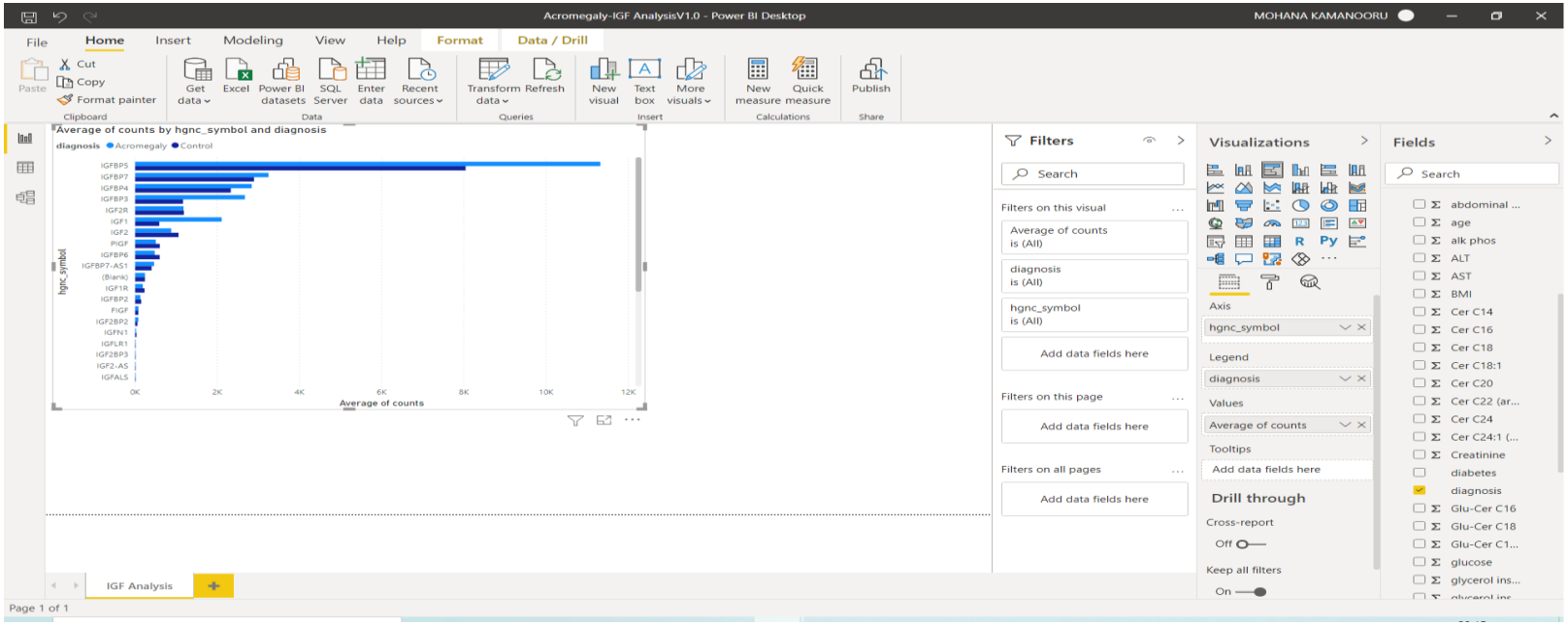
The snowflake schema with a fact-less fact table is from the data model mentioned in Business Intelligence Questions document. According to the business purposes and questions following relationships are analyzed.

1. IGF Analysis

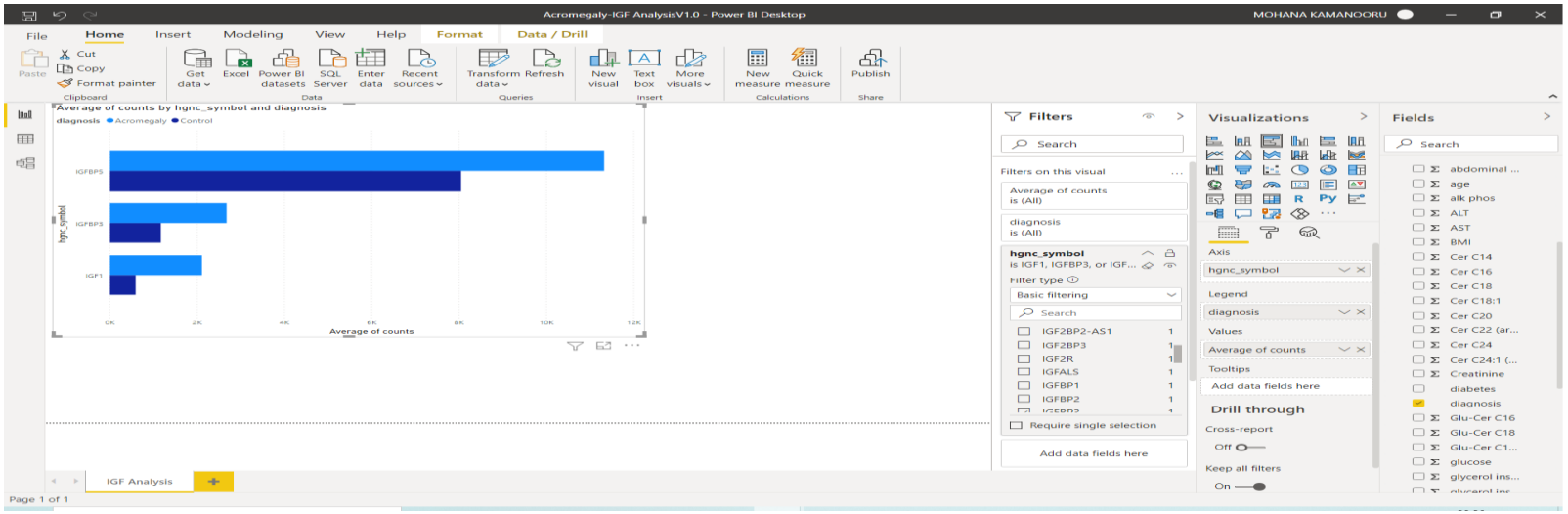
The primary symptom of Acromegaly is the secretion of excess growth hormone. To analyze this, plot graphs and evaluate IGF Complex("IGF1", "IGFBP3", "IGFBP5", "IGFALS"), IGF1 levels, IGF counts for both Acromegaly and Control patients.

1.1. IGF Complex

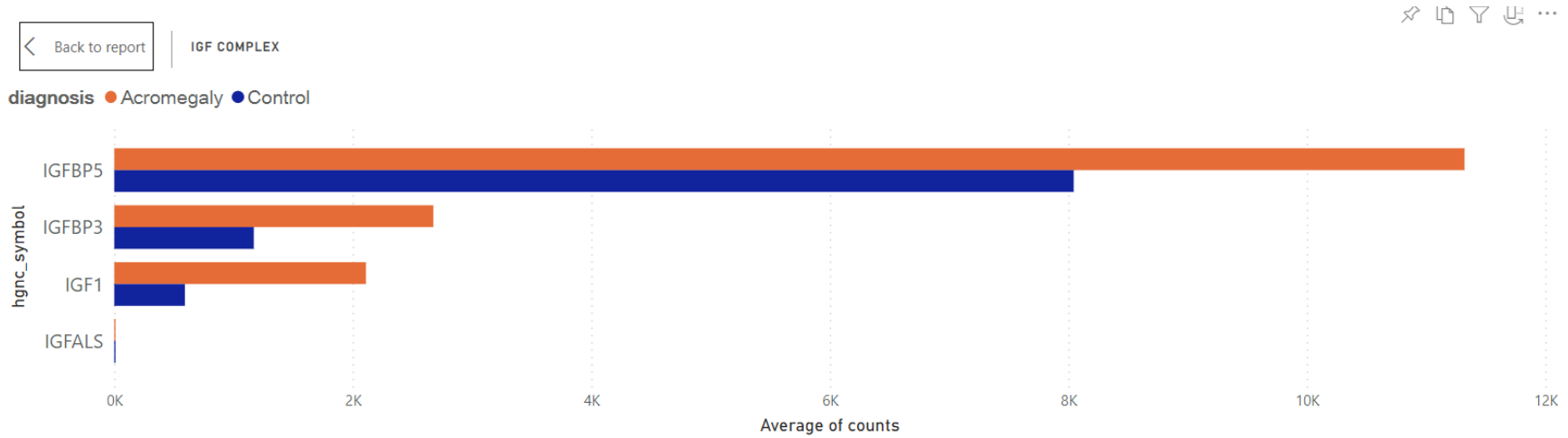
Plot Horizontal Bar graph between "**hgnc_symbol**" column from **Ensembl Gene Annotation** Table and Average of "**counts"** from **HTSEQ_Counts** Table. Analyze by "**diagnosis**" column as a legend from **patient_information** Table.



From Filters pane , check the boxes for "IGF1", "IGFBP3", "IGFBP5", "IGFALS" under basic filtering for hgnc_symbol.



And change the data color and font options from the Format section under the Visualization pane.

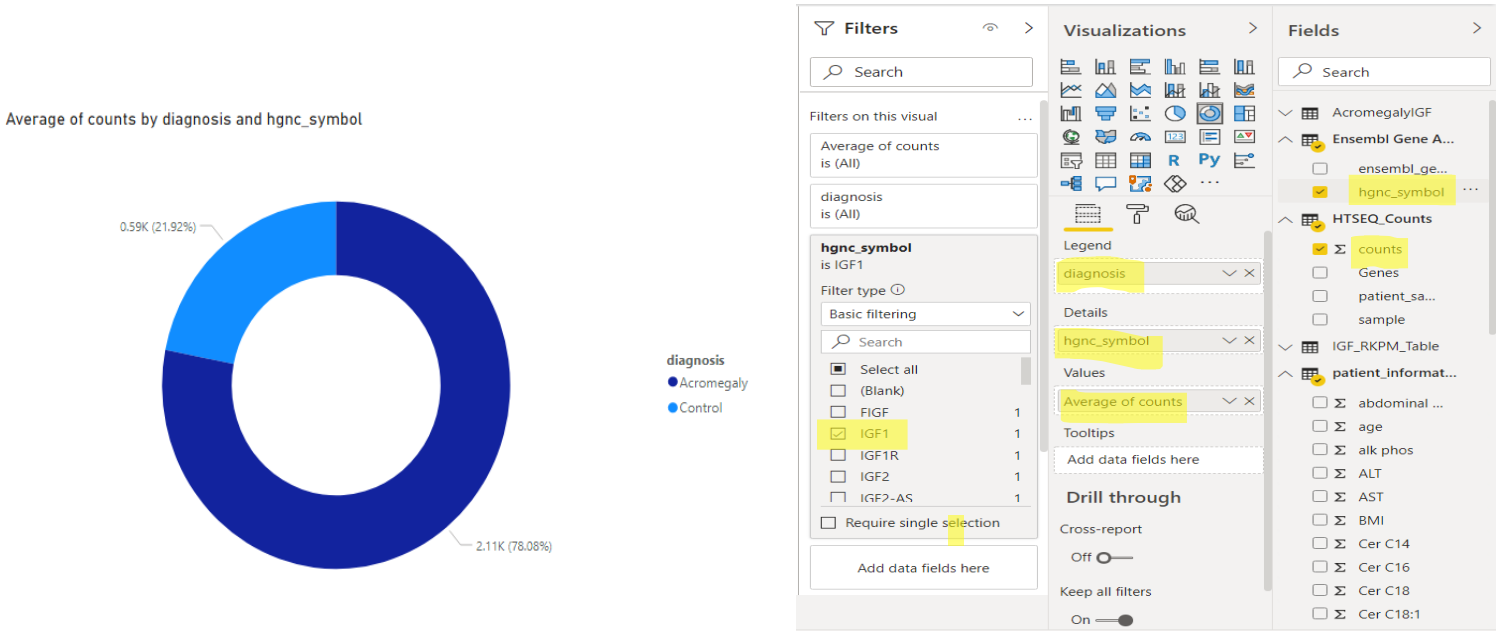


hgnc_symbol	Acromegaly	Control
IGFBP5	11321.57	8044.55
IGFBP3	2675.14	1169.73
IGF1	2110.00	592.36
IGFALS	2.86	2.09

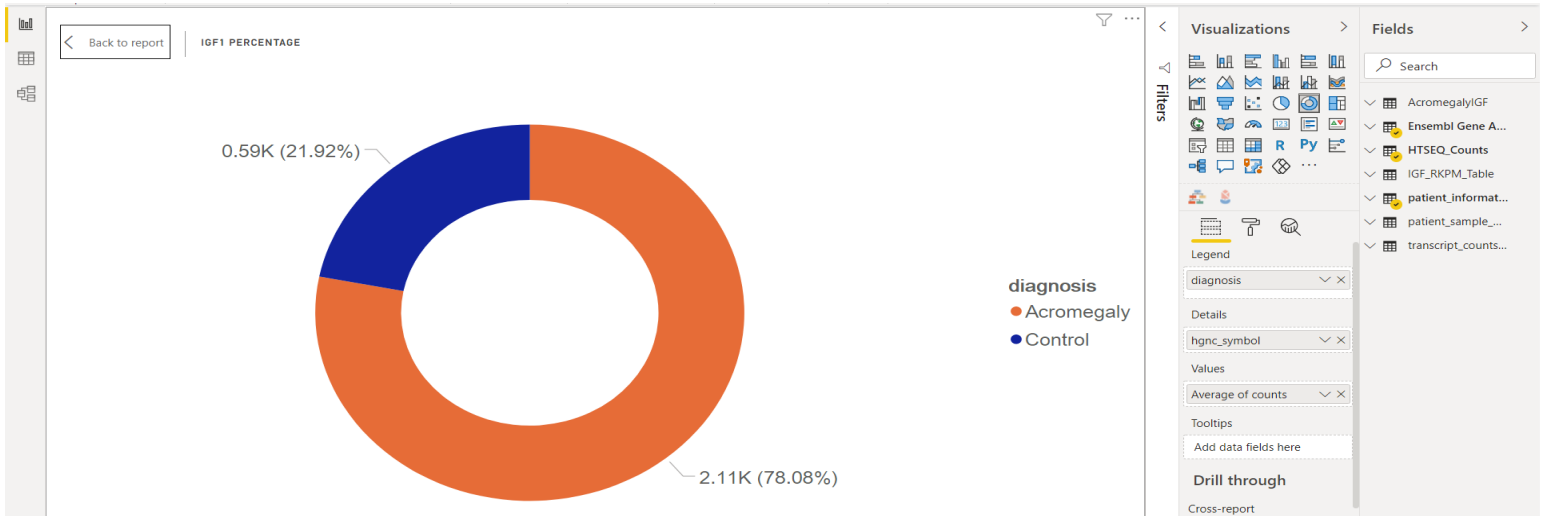
This horizontal bar graph will help to evaluate the IGF Complex function which involves selected genes and pictures the significant increase in IGF levels in Acromegaly patients compared to the control group.

1.2. IGF Levels

Plot donut graph to view the percentage difference between Acromegaly and Control group for IGF1 levels. Select donut graph from Visualization pane and select diagnosis column from patient_information Table as Legend. Average of counts from HTSEQ_Counts Table as values and hgnc_symbol as details and select IGF1 from the Filters pane under basic filtering for hgnc_symbol.

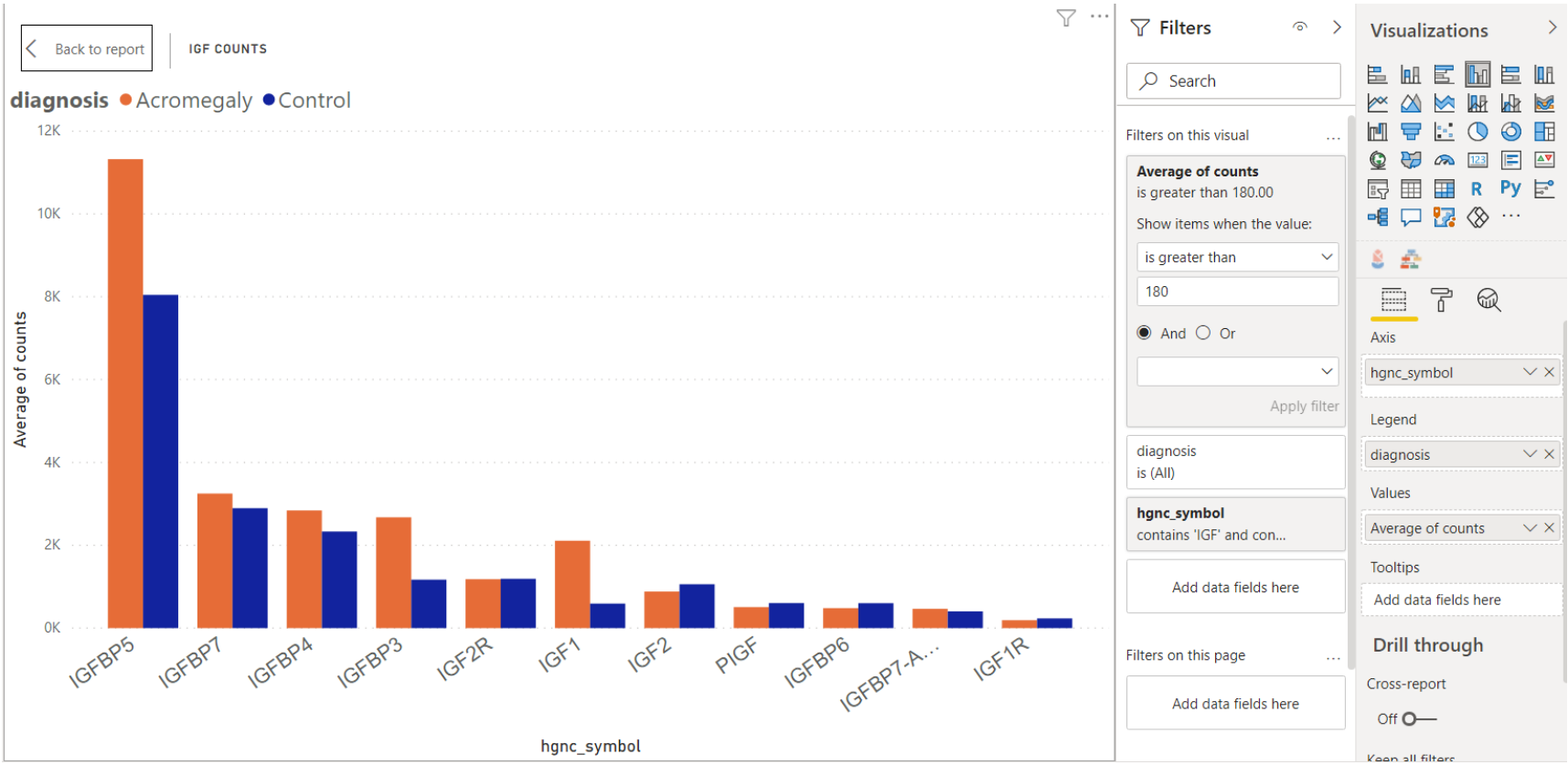


The plots look like the below after formatting the title, data color, and fonts from the Format section.



1.3. IGF Counts

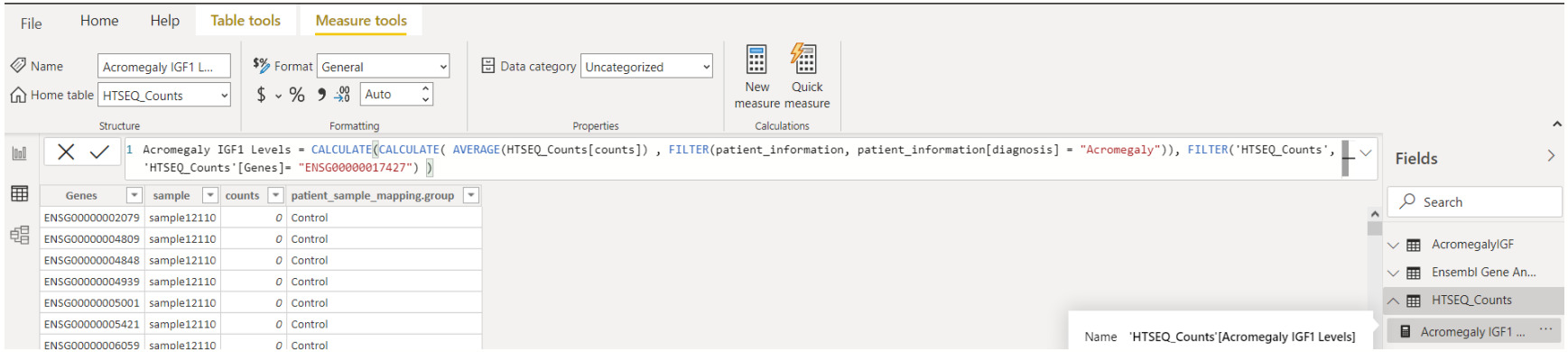
Plot horizontal bar graph to understand the difference in gene sequence counts. Use the same columns and details as mentioned in the IGF Complex plot from the Visualization pane. Select the values that contain "IGF" from the filters section in Filters pane for hgnc_symbol, and to see the highest counts filter Average of counts greater than 180. After formatting the title and axis fonts, the plots look like below.



1.4. IGF1 and IGFBP3 Measures

Create measures to visualize the numerical difference in the counts for Acromegaly and Control Patients' IGF counts and respective percentage differences. DAX formula is used to calculate the IGF levels.

IGF1 – Acromegaly

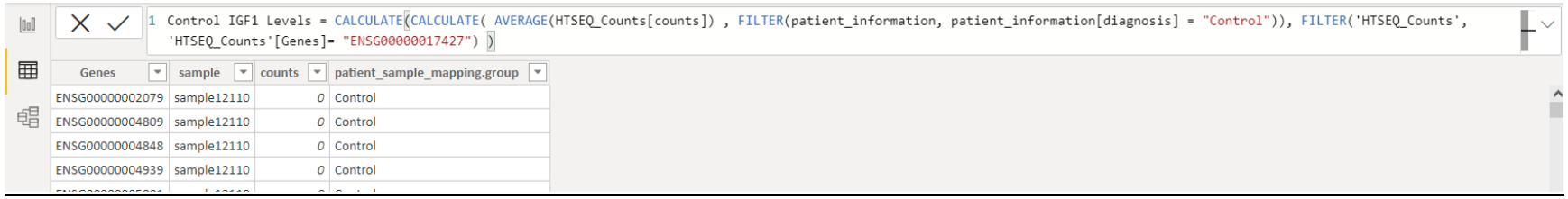



```
Acromegaly IGF1 Levels = CALCULATE(CALCULATE( AVERAGE(HTSEQ_Counts[counts]) , FILTER(patient_information,
patient_information[diagnosis] = "Acromegaly")), FILTER('HTSEQ_Counts','HTSEQ_Counts'[Genes]= "ENSG00000017427") )
```

1. FILTER('HTSEQ_Counts','HTSEQ_Counts'[Genes]= "*ENSG00000017427*")
   - Filters the HTSEQ_counts table for IGF1 , Gene Id for IGF1 is *ENSG00000017427*
2. FILTER(patient_information, patient_information[diagnosis] = "*Acromegaly*")
   - Filters the patient_information table for *Acromegaly* patients
3. AVERAGE(HTSEQ_Counts[counts])
   - Calculates the average of counts column from HTSEQ_Counts table.
4. CALCULATE(AVERAGE(….) , FILTER(….))
   - Calculates the (HTSEQ_Counts) counts average for Acromegaly patients.
5. CALCULATE(CALCULATE(….), FILTER(….))
   - Calculates the sequence counts for IGF1 genes.

Therefore, the DAX formula returns the average of sequence counts for Acromegaly patients for the IGF1 gene. Similarly, to calculate IGFBP3 genes count for both Acromegaly and Control patients use the following formulae.

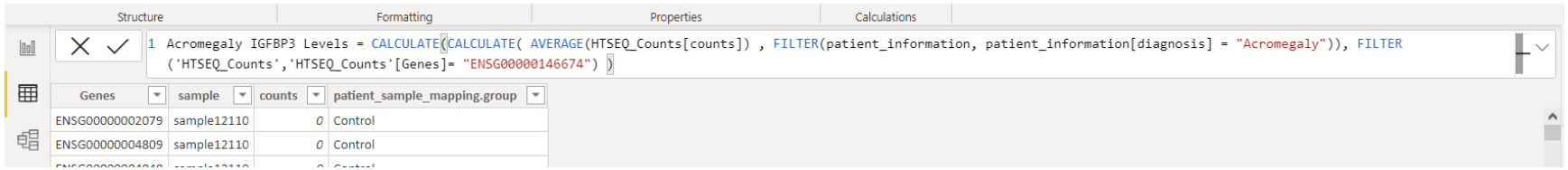
IGF1 – Control



Genes	sample	counts	patient_sample_mapping.group
ENSG00000002079	sample12110	0	Control
ENSG00000004809	sample12110	0	Control
ENSG00000004848	sample12110	0	Control
ENSG00000004939	sample12110	0	Control

Control IGF1 Levels = CALCULATE(CALCULATE(AVERAGE(HTSEQ_Counts[counts]) , FILTER(patient_information, patient_information[diagnosis] = "Control")), FILTER('HTSEQ_Counts','HTSEQ_Counts'[Genes]= "ENSG00000017427"))

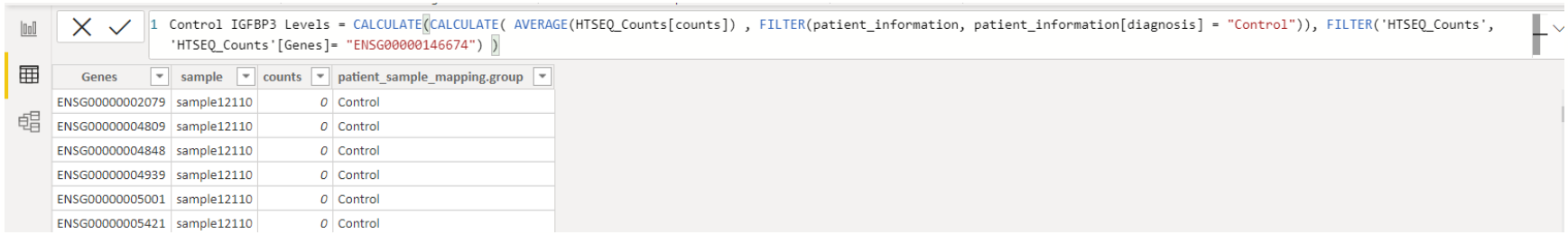
IGFBP3 – Acromegaly



Genes	sample	counts	patient_sample_mapping.group
ENSG00000002079	sample12110	0	Control
ENSG00000004809	sample12110	0	Control

Acromegaly IGFBP3 Levels = CALCULATE(CALCULATE(AVERAGE(HTSEQ_Counts[counts]) , FILTER(patient_information, patient_information[diagnosis] = "Acromegaly")), FILTER('HTSEQ_Counts','HTSEQ_Counts'[Genes]= "ENSG00000146674"))

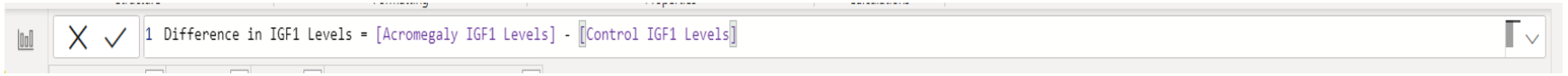
IGFBP3 - Control

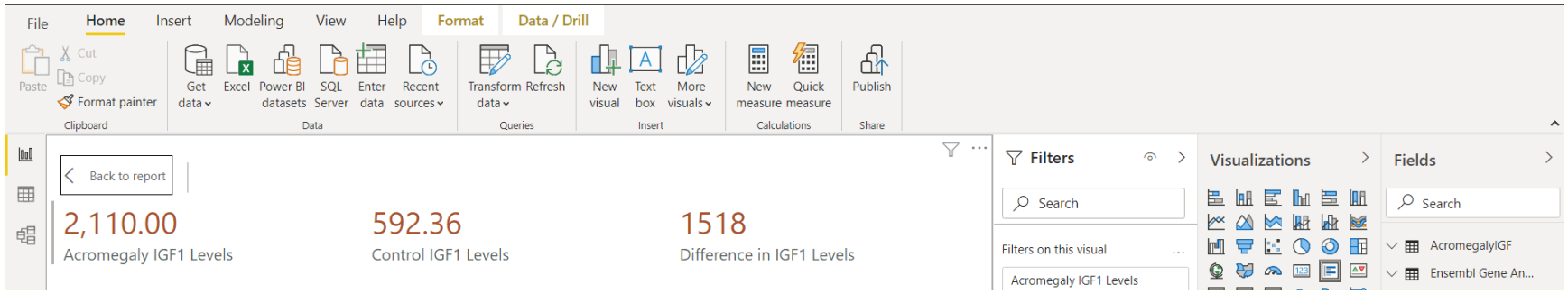


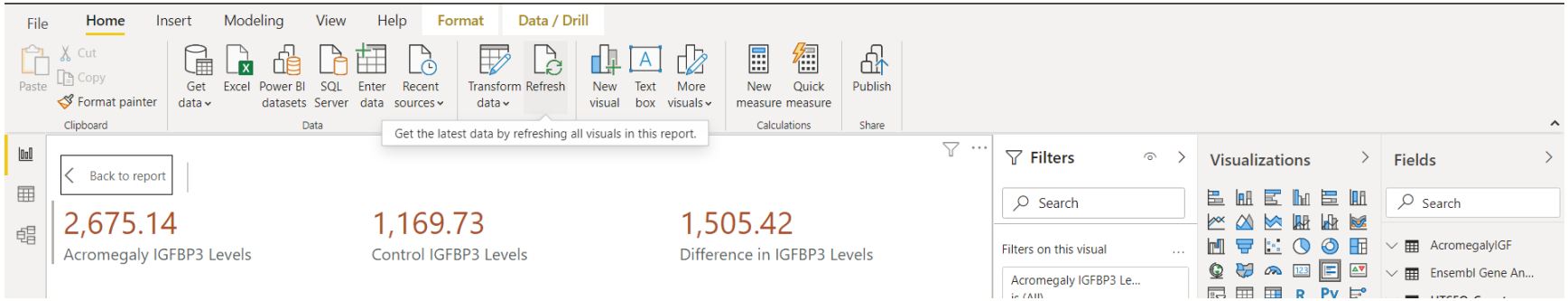
Genes	sample	counts	patient_sample_mapping.group
ENSG00000002079	sample12110	0	Control
ENSG00000004809	sample12110	0	Control
ENSG00000004848	sample12110	0	Control
ENSG00000004939	sample12110	0	Control
ENSG00000005001	sample12110	0	Control
ENSG00000005421	sample12110	0	Control

```
Control IGFBP3 Levels = CALCULATE(CALCULATE( AVERAGE(HTSEQ_Counts[counts]) , FILTER(patient_information, patient_information[diagnosis] = "Control")), FILTER('HTSEQ_Counts','HTSEQ_Counts'[Genes]= "ENSG00000146674") )
```

Difference in IGF1 and IGFBP3 levels







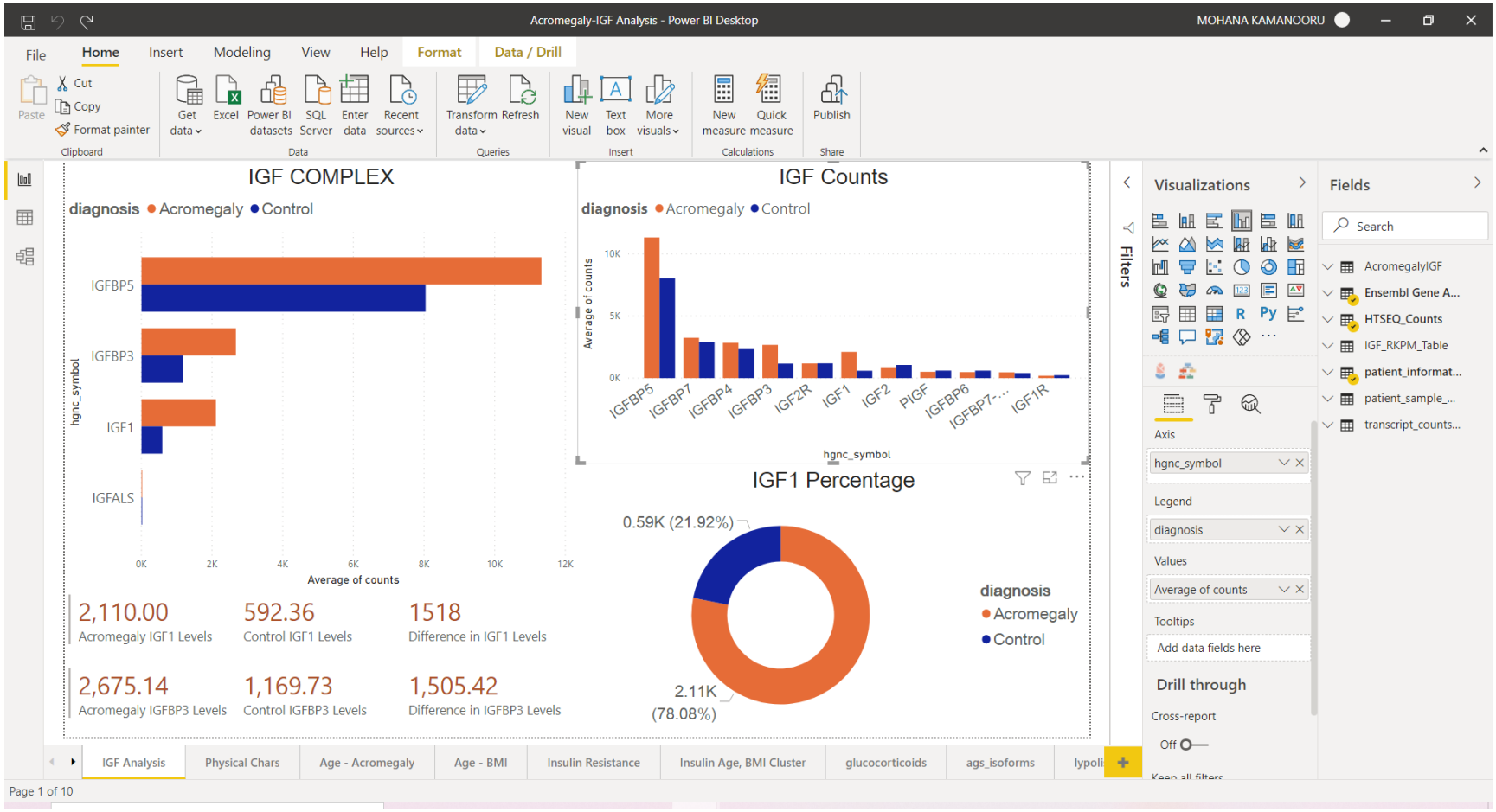


Figure 2 IGF Analysis Plots

1.5. Analytics

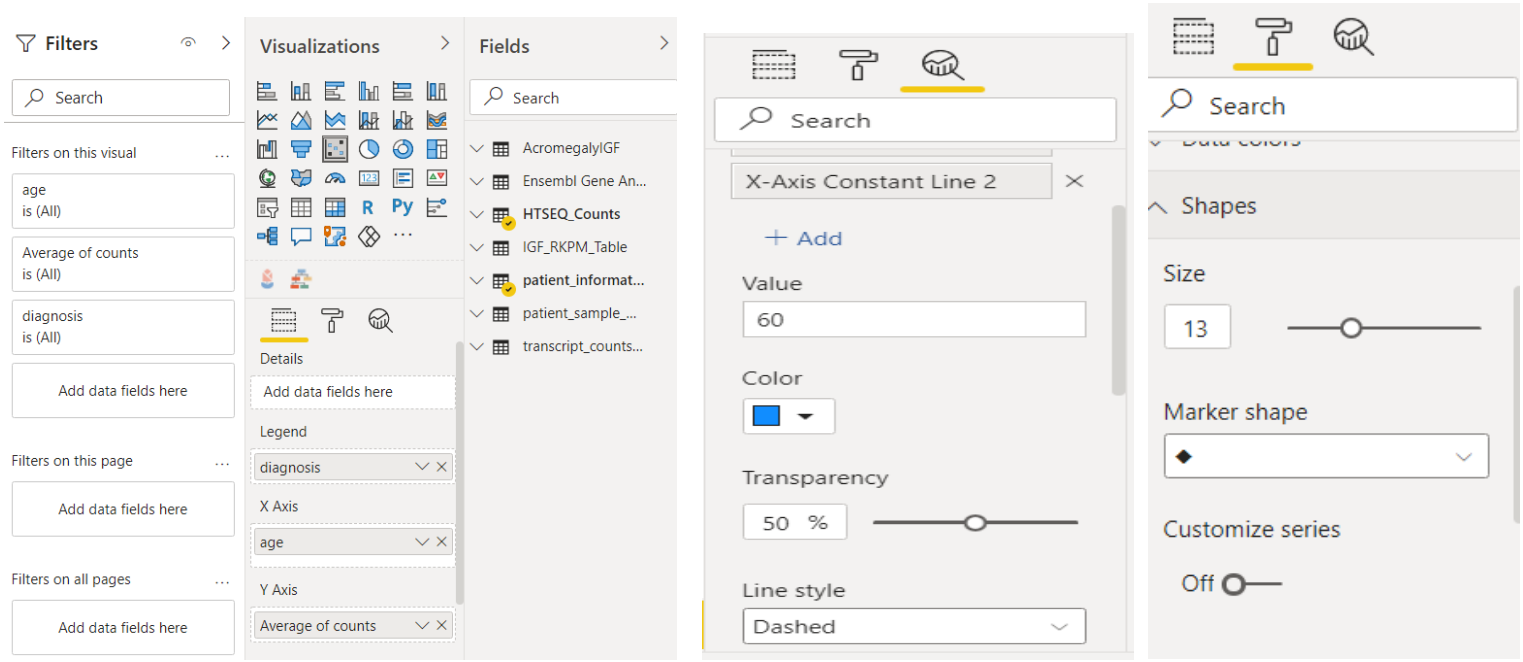
From the plots, IGF Complex and IGF Count, it is evident that the levels of Growth Hormone are very high in Acromegaly patients compared to the Control group. The percentage of IGF1 is 78% in Acromegaly and 22% in Controls.

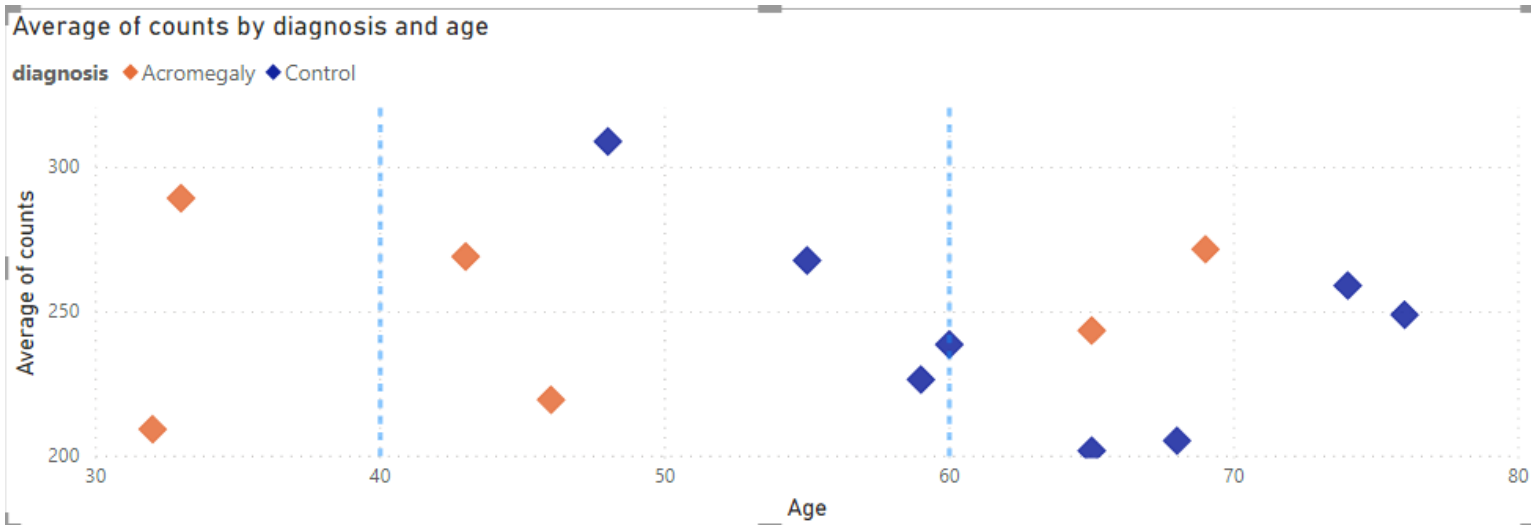
1.6. Findings

From analytics and observations, If the pituitary gland is affected by the tumor and if it releases high levels of growth hormone then the probability of having Acromegaly could be quite high in the patient.

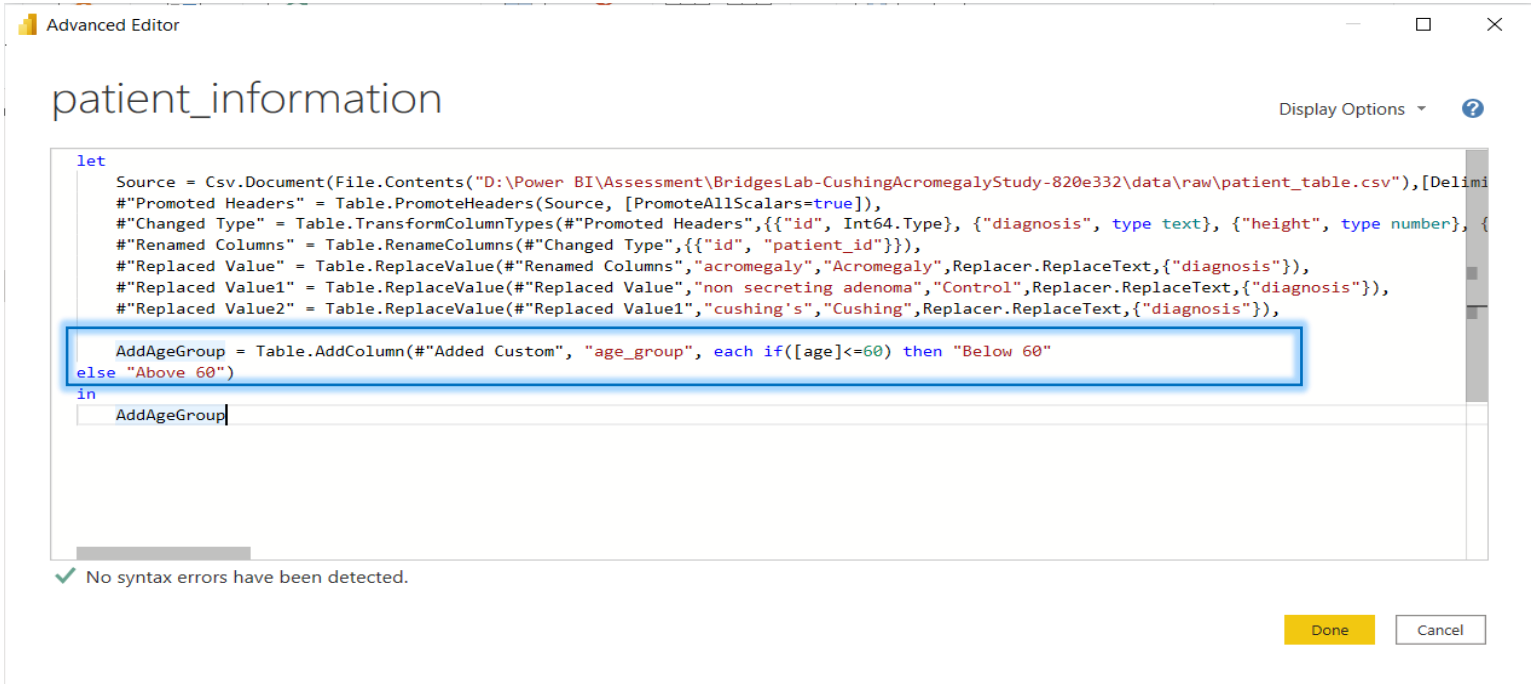
2. Age Analysis

To analyze if Acromegaly is age affected or not, plot the scatter graph between Age of all patients and the average of HTSEQ_Counts. From the Analytics tab add two constant lines on the x-axis, to view age in three blocks.

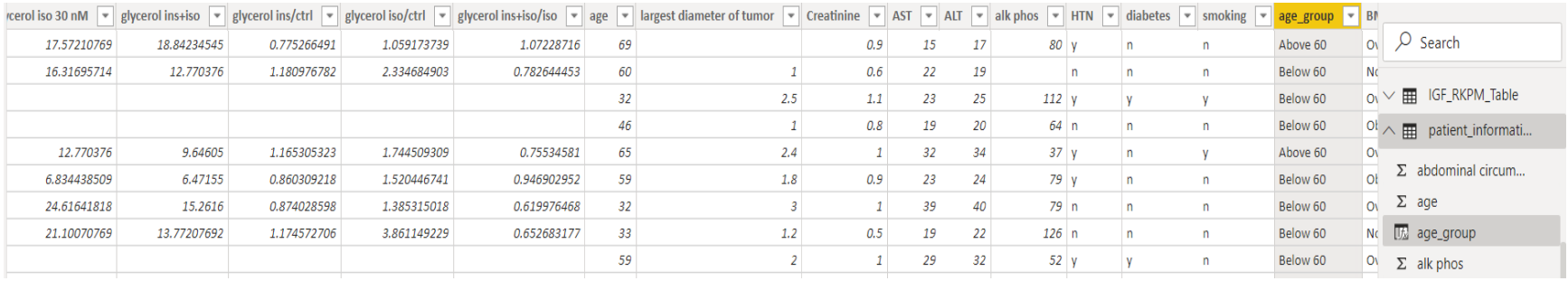




Since the data is not very clear in the plot to draw more information, divide the Age column into two blocks Above 60 and Below 60 using M language. Click on the Transform Data tab and select the patient_information table from the Power Query Editor. Click on Advanced Editor.



Click Done, Now the new column age_group is created in the table. Click on Apply and Close.



ycerol iso 30 nM	glycerol ins+iso	glycerol ins/ctrl	glycerol iso/ctrl	glycerol ins+iso/iso	age	largest diameter of tumor	Creatinine	AST	ALT	alk phos	HTN	diabetes	smoking	age_group	BM
17.57210769	18.84234545	0.775266491	1.059173739	1.07228716	69		0.9	15	17	80	y	n	n	Above 60	Ov
16.31695714	12.770376	1.180976782	2.334684903	0.782644453	60	1	0.6	22	19		n	n	n	Below 60	No
					32	2.5	1.1	23	25	112	y	y	y	Below 60	Ov
					46	1	0.8	19	20	64	n	n	n	Below 60	Ob
12.770376	9.64605	1.165305323	1.744509309	0.75534581	65	2.4	1	32	34	37	y	n	y	Above 60	Ov
6.834438509	6.47155	0.860309218	1.520446741	0.946902952	59	1.8	0.9	23	24	79	y	n	n	Below 60	Ob
24.61641818	15.2616	0.874028598	1.385315018	0.619976468	32	3	1	39	40	79	n	n	n	Below 60	Ov
21.10070769	13.77207692	1.174572706	3.861149229	0.652683177	33	1.2	0.5	19	22	126	n	n	n	Below 60	No
					59	2	1	29	32	52	y	y	n	Below 60	Ov

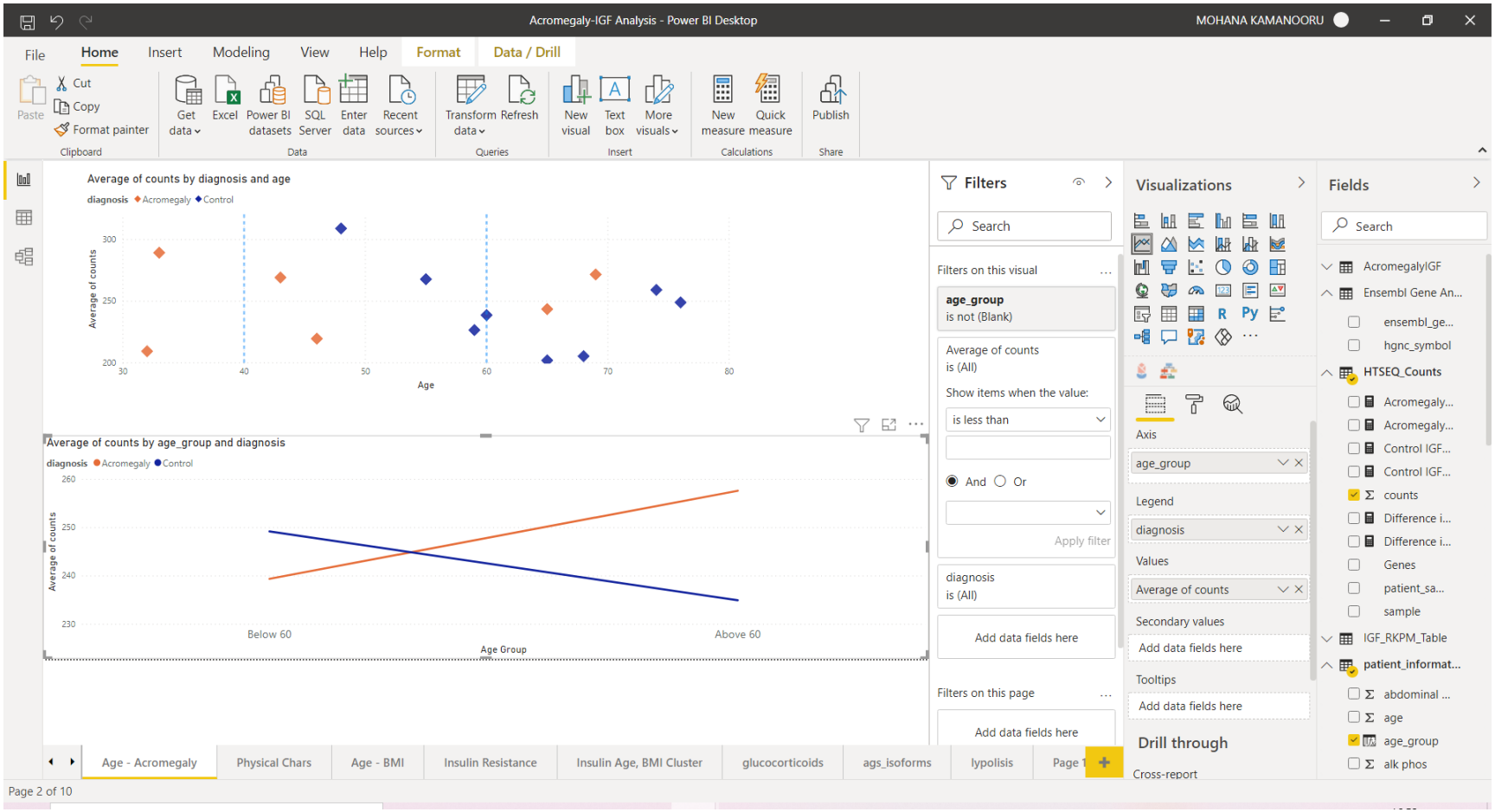


Figure 3 Age Analysis Plots

2.1. Analytics

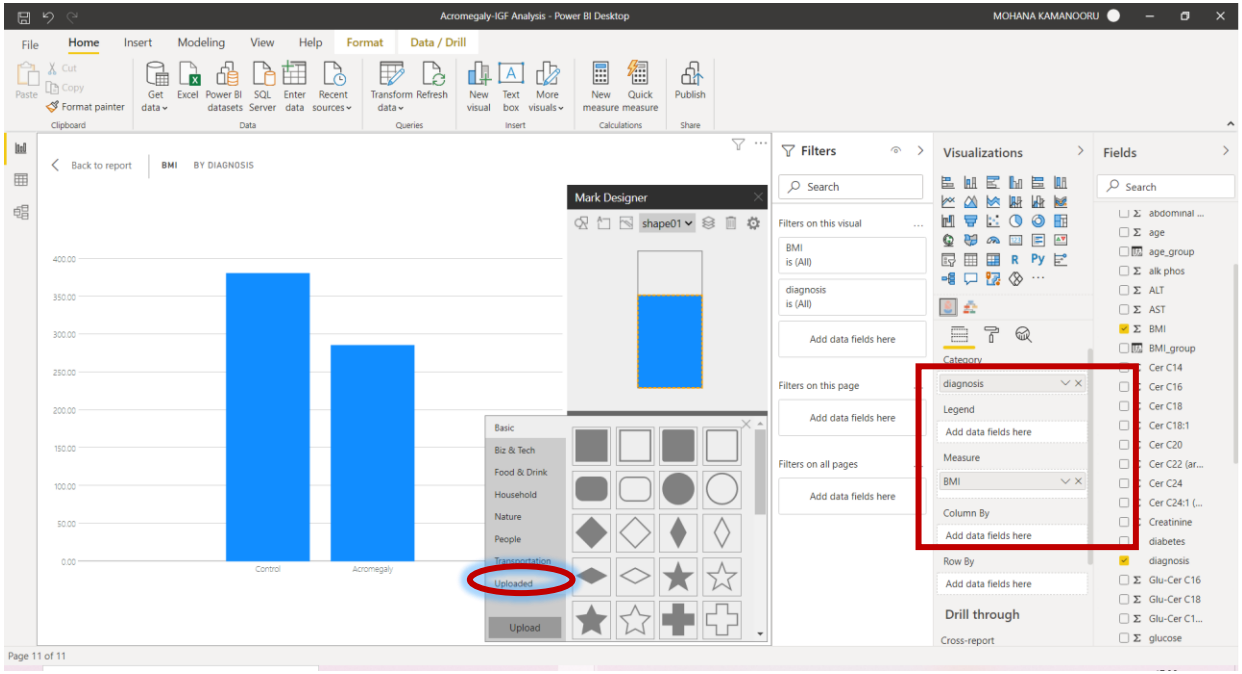
From the plot, the average counts for Acromegaly patients tend to be higher, there are three patients above 60 years and 6 patients below 60 years among a total of nine Acromegaly patients. This implies the elevation in gene levels occurs as the patient gets older and gets adapted to the illness.

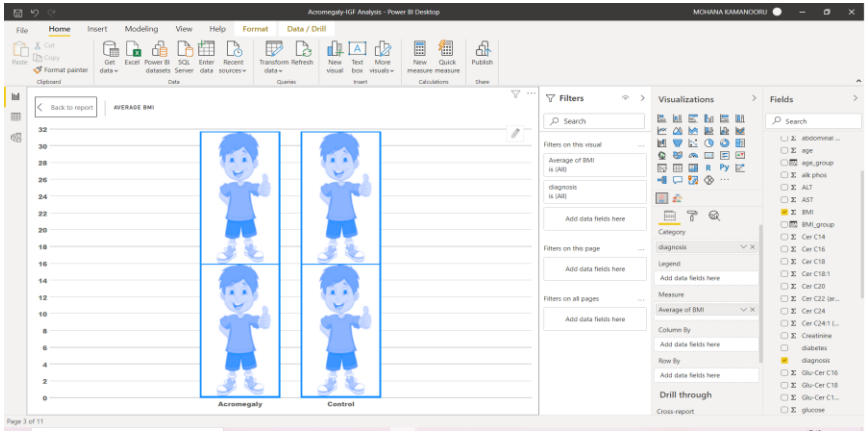
2.2. Findings

Considering the sample size of this secondary research, the difference in counts is observed to be marginal. There might be slightly considerable effects of age on Acromegaly but cannot be decided if this could be appropriate for a larger sample size.

3. Physical Characteristics Analysis

Draw a vertical bar plot with infographics for both groups of patients as shown in screenshots below. Select BMI from Patient_information table and diagnosis column as a legend. Select the bar plot and click on infographics, choose to upload a new custom image, and select it.





Repeat the same steps for different columns (height, weight, and abdominal circumference) in the patient_information table. The plots are shown below in the screenshot.

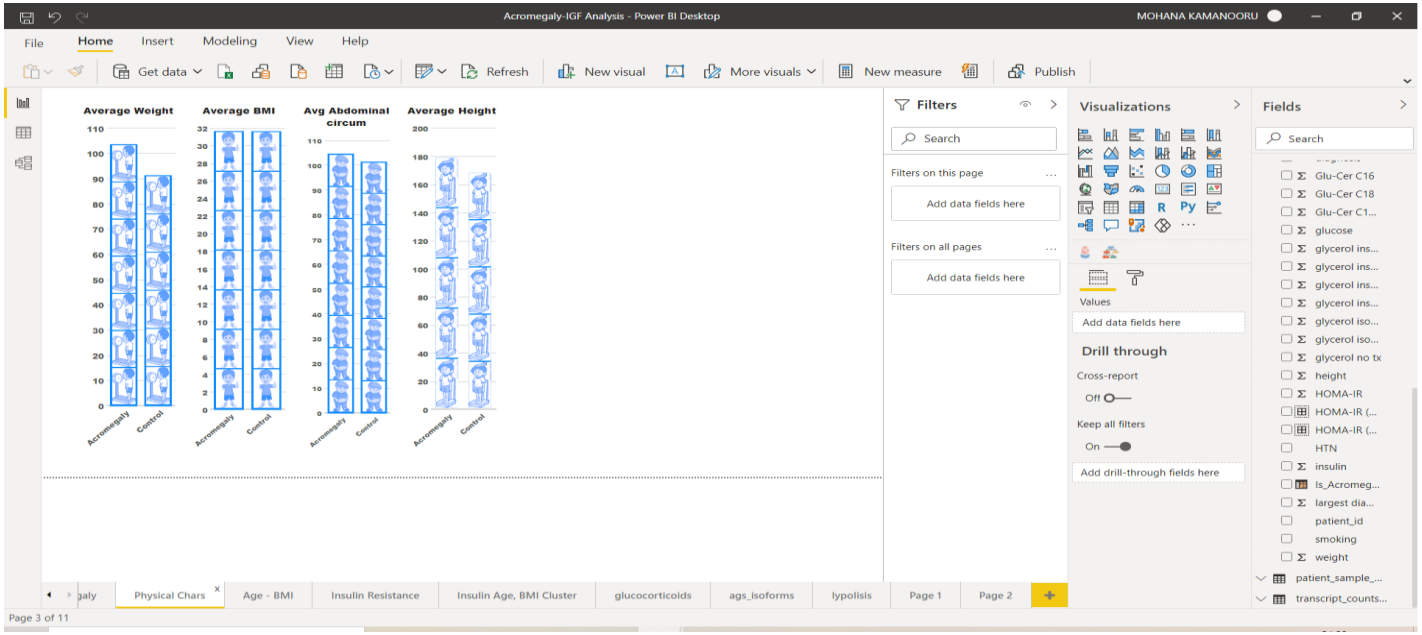


Figure 4 Physical Chars Analysis Plots

3.1. Analytics

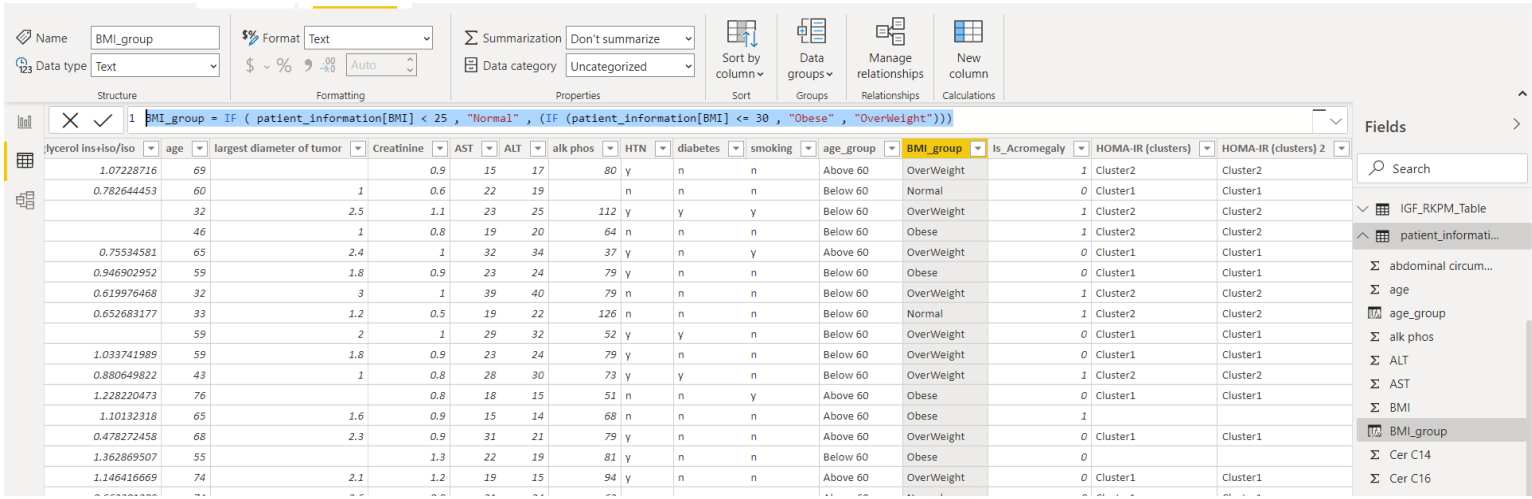
No significant differences are found in physical characteristics between patient groups while analyzing BMI and abdominal circumferences. Except the Acromegaly patients were taller and weighed more with an average difference of 10kgs.

3.2. Findings

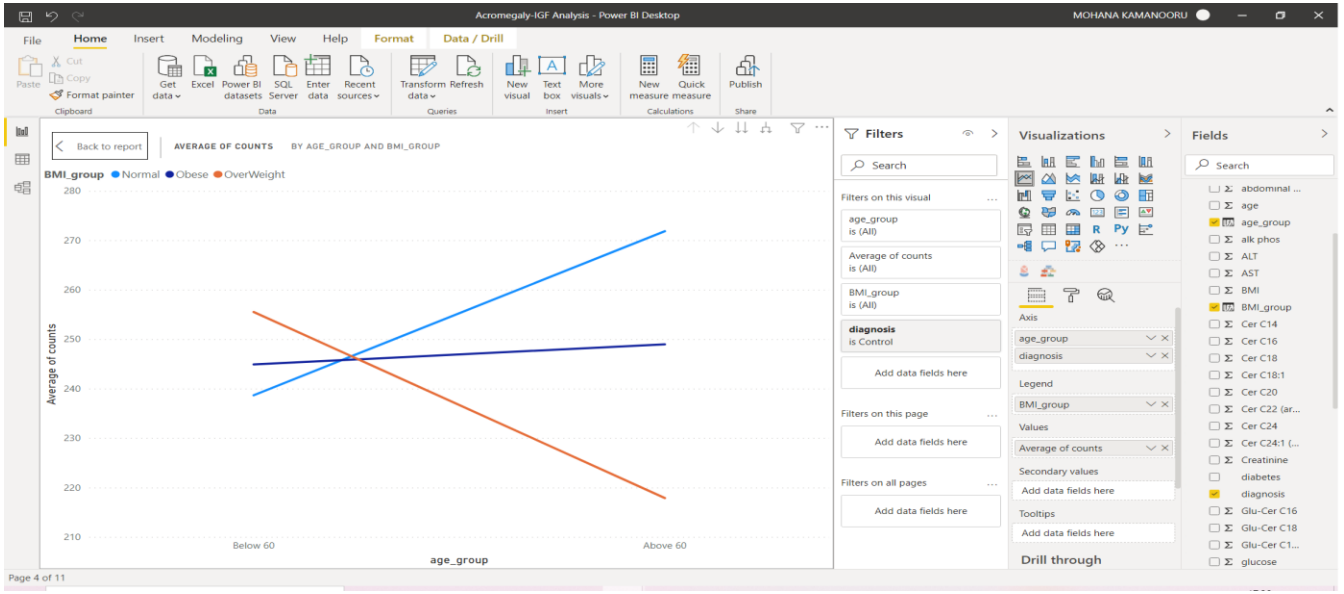
There are many studies and research papers available on the physical characteristics of Acromegaly patients and are mostly related to the face, hands, tongue, and feet. Since there is no data recorded in the current dataset, these characteristics could not be analyzed in this research. The physical characteristics (height, weight, BMI, and abdominal circumferences) do not tend to change in considerable amounts. This can be ignored for the sample size and magnitude of the differences.

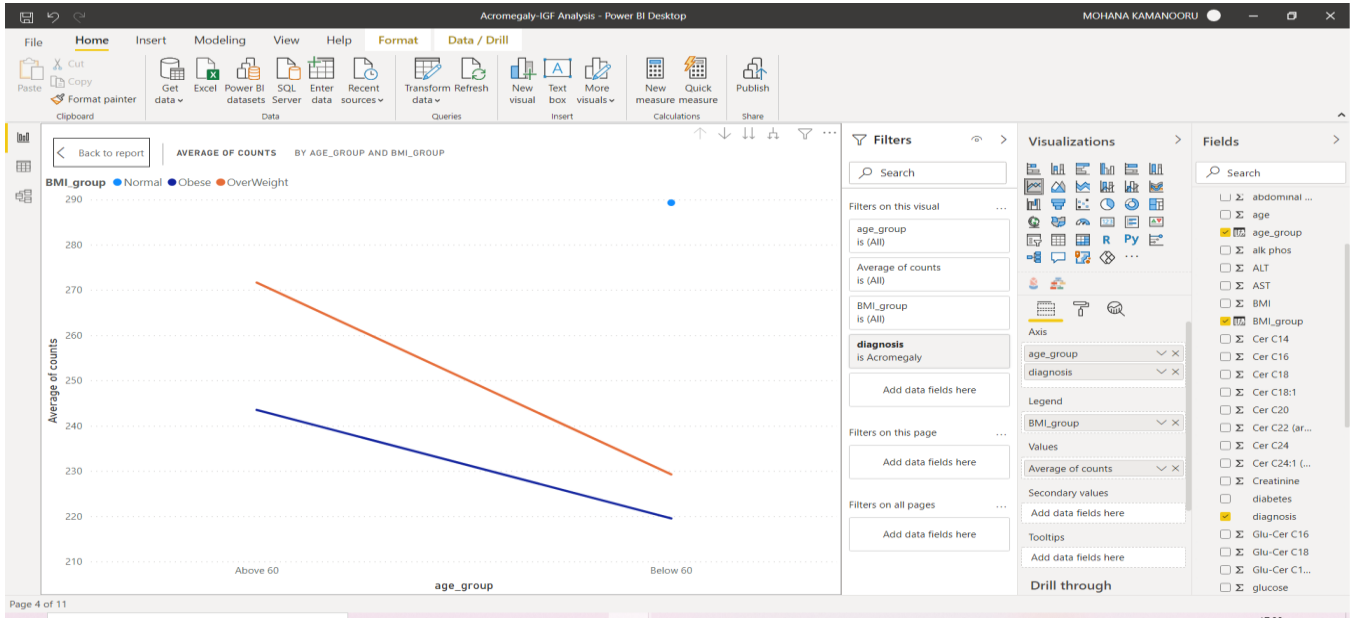
4. BMI Analysis

Plot the line graph between Age, BMI, and the counts. To analyze the relation for both patient groups if any. Firstly, categorize the BMI column into three blocks Normal (< 25), Obese (25 to 30), Overweight (>30) using the DAX formula as below.



lycerol ins+iso/iso	age	largest diameter of tumor	Creatinine	AST	ALT	alk phos	HTN	diabetes	smoking	age_group	BMI_group	Is_Acromegaly	HOMA-IR (clusters)	HOMA-IR (clusters) 2
1.07228716	69		0.9	15	17	80	y	n	n	Above 60	OverWeight	1	Cluster2	Cluster2
0.782644453	60	1	0.6	22	19		n	n	n	Below 60	Normal	0	Cluster1	Cluster1
	32	2.5	1.1	23	25	112	y	y	y	Below 60	OverWeight	1	Cluster2	Cluster2
	46	1	0.8	19	20	64	n	n	n	Below 60	Obese	1	Cluster2	Cluster2
0.75534581	65	2.4	1	32	34	37	y	n	y	Above 60	OverWeight	0	Cluster1	Cluster1
0.946902952	59	1.8	0.9	23	24	79	y	n	n	Below 60	Obese	0	Cluster1	Cluster1
0.619976468	32	3	1	39	40	79	n	n	n	Below 60	OverWeight	1	Cluster2	Cluster2
0.652683177	33	1.2	0.5	19	22	126	n	n	n	Below 60	Normal	1	Cluster2	Cluster2
	59	2	1	29	32	52	y	y	n	Below 60	OverWeight	0	Cluster1	Cluster1
1.033741989	59	1.8	0.9	23	24	79	y	n	n	Below 60	OverWeight	0	Cluster1	Cluster1
0.880649822	43	1	0.8	28	30	73	y	y	n	Below 60	OverWeight	1	Cluster2	Cluster2
1.228220473	76		0.8	18	15	51	n	n	y	Above 60	Obese	0	Cluster1	Cluster1
1.10132318	65	1.6	0.9	15	14	68	n	n	n	Above 60	Obese	1		
0.478272458	68	2.3	0.9	31	21	79	y	n	n	Above 60	OverWeight	0	Cluster1	Cluster1
1.362869507	55		1.3	22	19	81	y	n	n	Below 60	Obese	0		
1.146416669	74	2.1	1.2	19	15	94	y	n	n	Above 60	OverWeight	0	Cluster1	Cluster1


Acromegaly-IGF Analysis - Power BI Desktop
MOHANA KAMANOORU
File
Home
Insert
Modeling
View
Help
Format
Data / Drill
Back to report
AVERAGE OF COUNTS
BY AGE_GROUP AND BMI_GROUP
BMI_group Normal Obese OverWeight
Average of counts
age_group
Below 60
Above 60
Filters
Search
Filters on this visual
age_group is (All)
Average of counts is (All)
BMI_group is (All)
diagnosis is Control
Add data fields here
Filters on this page
Filters on all pages
Visualizations
Axis
Legend
Values
Secondary values
Tooltips
Drill through
Fields
abdominal ...
age
age_group
alk phos
ALT
AST
BMI
BMI_group
Cer C14
Cer C16
Cer C18
Cer C18:1
Cer C20
Cer C22 (ar...
Cer C24
Creatinine
diabetes
diagnosis
Glu-Cer C16
Glu-Cer C18
glucose
Page 4 of 11


Acromegaly-IGF Analysis - Power BI Desktop
MOHANA KAMANOORU
File Home Insert Modeling View Help Format Data / Drill
Back to report
AVERAGE OF COUNTS BY AGE_GROUP AND BMI_GROUP
BMI_group Normal Obese OverWeight
Average of counts
290 280 270 260 250 240 230 220 210
Above 60 Below 60
age_group
Page 4 of 11
Filters
Search
Filters on this visual
age_group is (All)
Average of counts is (All)
BMI_group is (All)
diagnosis is Acromegaly
Add data fields here
Filters on this page
Add data fields here
Filters on all pages
Add data fields here
Visualizations
Axis
age_group
diagnosis
Legend
BMI_group
Values
Average of counts
Secondary values
Add data fields here
Tooltips
Add data fields here
Drill through
Fields
Search
abdominal ...
age
age_group
alk phos
ALT
AST
BMI
BMI_group
Cer C14
Cer C16
Cer C18
Cer C18:1
Cer C20
Cer C22 (ar...
Cer C24
Cer C24:1 (...
Creatinine
diabetes
diagnosis
Glu-Cer C16
Glu-Cer C18
Glu-Cer C1...
glucose

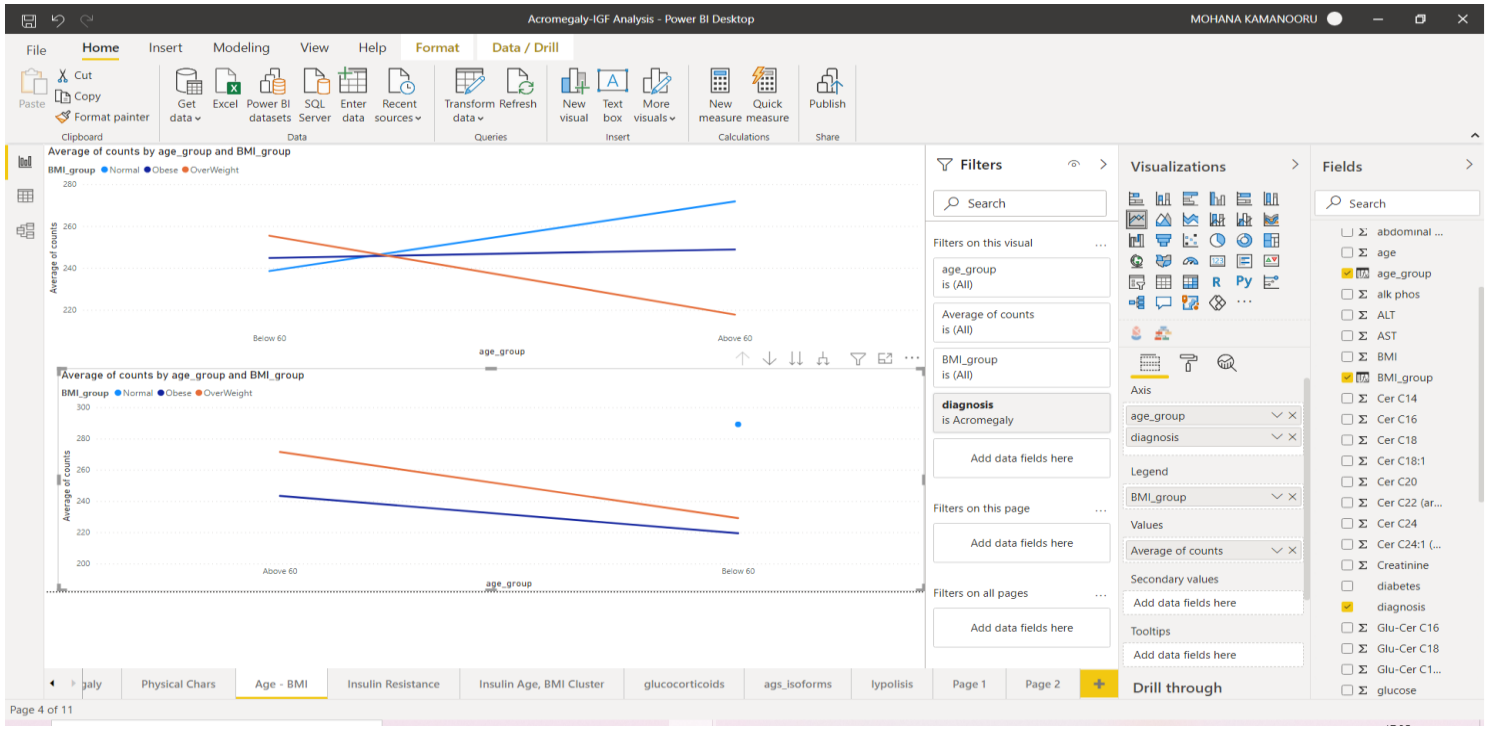


Figure 5 Age and BMI Analysis Plots

4.1. Analytics

No significant trends, patterns, or differences could be observed that could help analyze the illness.

4.2. Findings

No major observations.

5. Insulin (HOMA-IR) Analysis

To comprehend the insulin resistance and insulin sensitivity in both patient groups. Plot the below bar and line graphs between Age, BMI, and HOMA-IR values from patient_information Table.

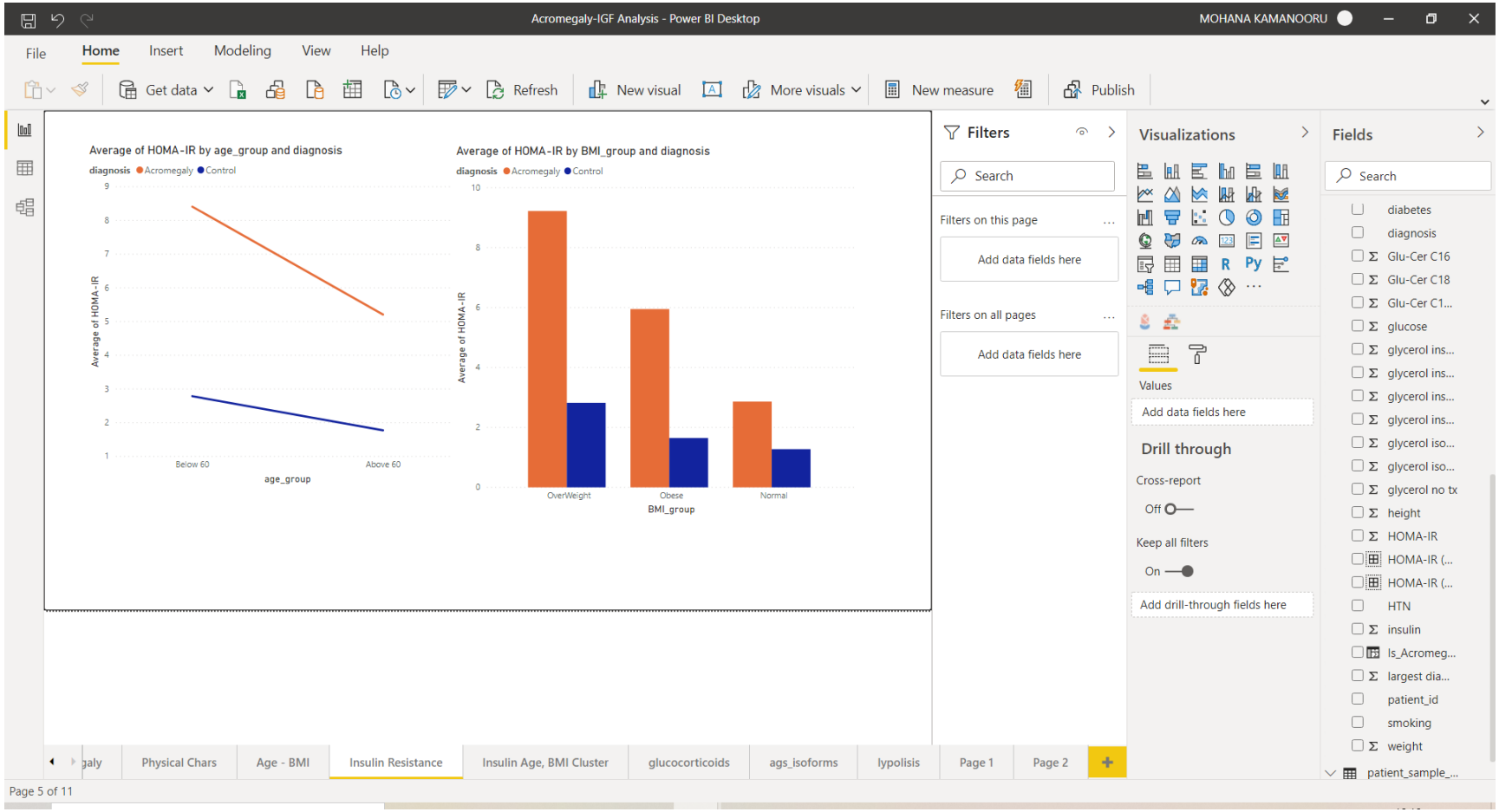
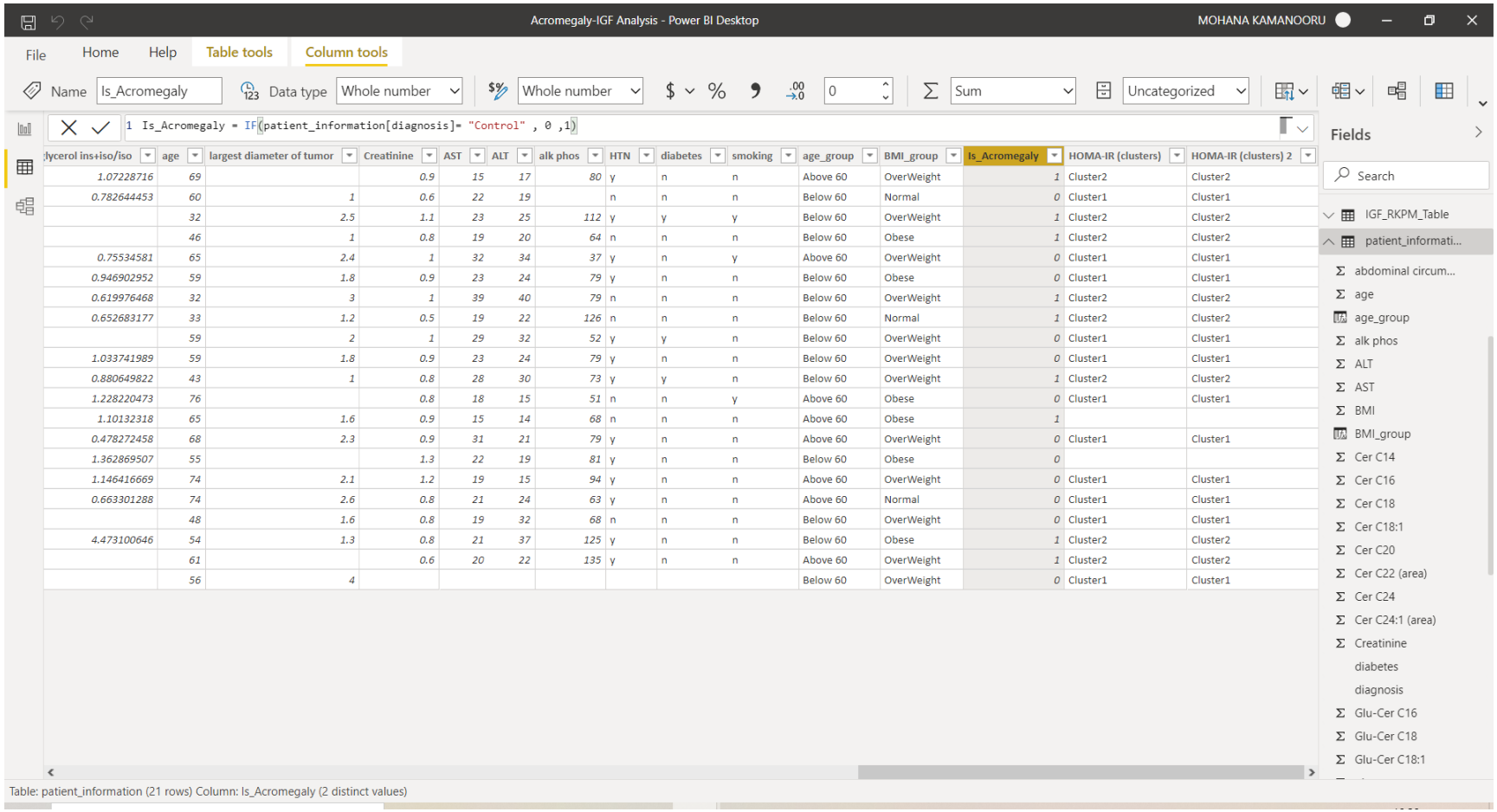


Figure 6 Insulin Resistance Plots

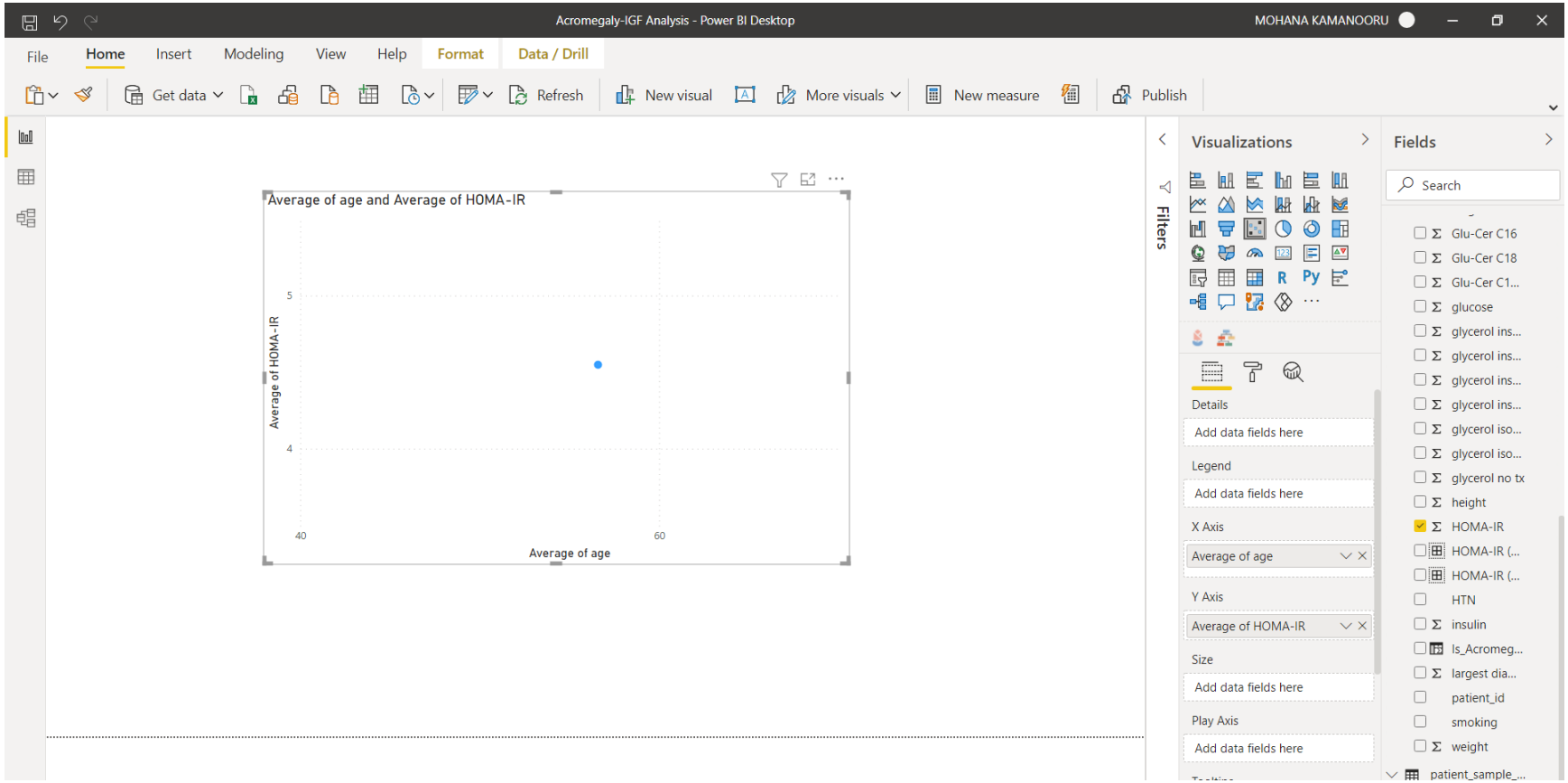
Acromegaly patients have higher insulin resistance and lower insulin sensitivity. Analyzing the plots, the insulin resistance is higher in patients below 60 years, and patients with higher BMI have higher insulin resistance.

Identifying Acromegaly patients with HOMA-IR, Age, and BMI using Artificial Intelligence (Clustering Algorithm) in power BI. Add a new column in Patient_information table “is_Acromegaly”, which holds a numeric value, 1 if True, 0 if False using DAX formula as shown in the screenshot below.

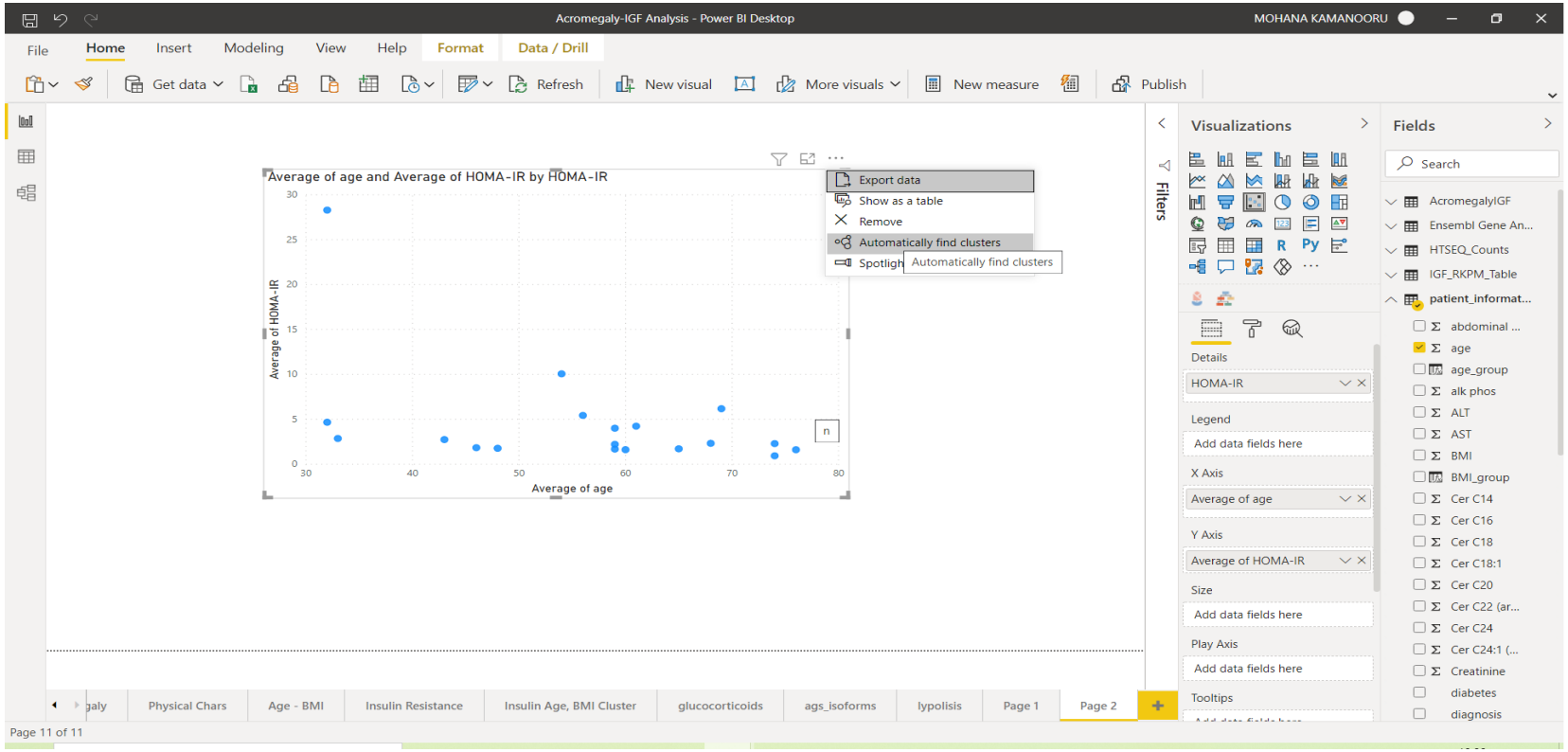


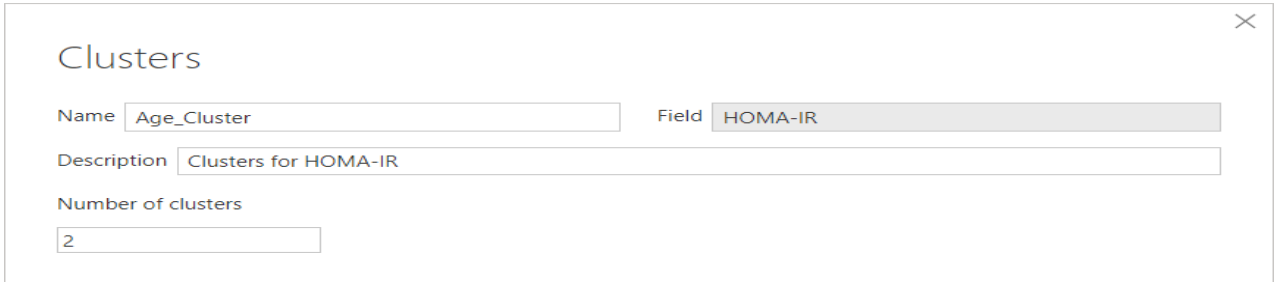
lycerol ins+iso/iso	age	largest diameter of tumor	Creatinine	AST	ALT	alk phos	HTN	diabetes	smoking	age_group	BMI_group	Is_Acromegaly	HOMA-IR (clusters)	HOMA-IR (clusters) 2
1.07228716	69		0.9	15	17	80	y	n	n	Above 60	OverWeight	1	Cluster2	Cluster2
0.782644453	60	1	0.6	22	19		n	n	n	Below 60	Normal	0	Cluster1	Cluster1
	32	2.5	1.1	23	25	112	y	y	y	Below 60	OverWeight	1	Cluster2	Cluster2
	46	1	0.8	19	20	64	n	n	n	Below 60	Obese	1	Cluster2	Cluster2
0.75534581	65	2.4	1	32	34	37	y	n	y	Above 60	OverWeight	0	Cluster1	Cluster1
0.946902952	59	1.8	0.9	23	24	79	y	n	n	Below 60	Obese	0	Cluster1	Cluster1
0.619976468	32	3	1	39	40	79	n	n	n	Below 60	OverWeight	1	Cluster2	Cluster2
0.652683177	33	1.2	0.5	19	22	126	n	n	n	Below 60	Normal	1	Cluster2	Cluster2
	59	2	1	29	32	52	y	y	n	Below 60	OverWeight	0	Cluster1	Cluster1
1.033741989	59	1.8	0.9	23	24	79	y	n	n	Below 60	OverWeight	0	Cluster1	Cluster1
0.880649822	43	1	0.8	28	30	73	y	y	n	Below 60	OverWeight	1	Cluster2	Cluster2
1.228220473	76		0.8	18	15	51	n	n	y	Above 60	Obese	0	Cluster1	Cluster1
1.10132318	65	1.6	0.9	15	14	68	n	n	n	Above 60	Obese	1		
0.478272458	68	2.3	0.9	31	21	79	y	n	n	Above 60	OverWeight	0	Cluster1	Cluster1
1.362869507	55		1.3	22	19	81	y	n	n	Below 60	Obese	0		
1.146416669	74	2.1	1.2	19	15	94	y	n	n	Above 60	OverWeight	0	Cluster1	Cluster1
0.663301288	74	2.6	0.8	21	24	63	y	n	n	Above 60	Normal	0	Cluster1	Cluster1
	48	1.6	0.8	19	32	68	n	n	n	Below 60	OverWeight	0	Cluster1	Cluster1
4.473100646	54	1.3	0.8	21	37	125	y	n	n	Below 60	Obese	1	Cluster2	Cluster2
	61		0.6	20	22	135	y	n	n	Above 60	OverWeight	1	Cluster2	Cluster2
	56	4								Below 60	OverWeight	0	Cluster1	Cluster1

Plot a scatter plot with the Average age column and Average of HOMA-IR column in patient_information Table.

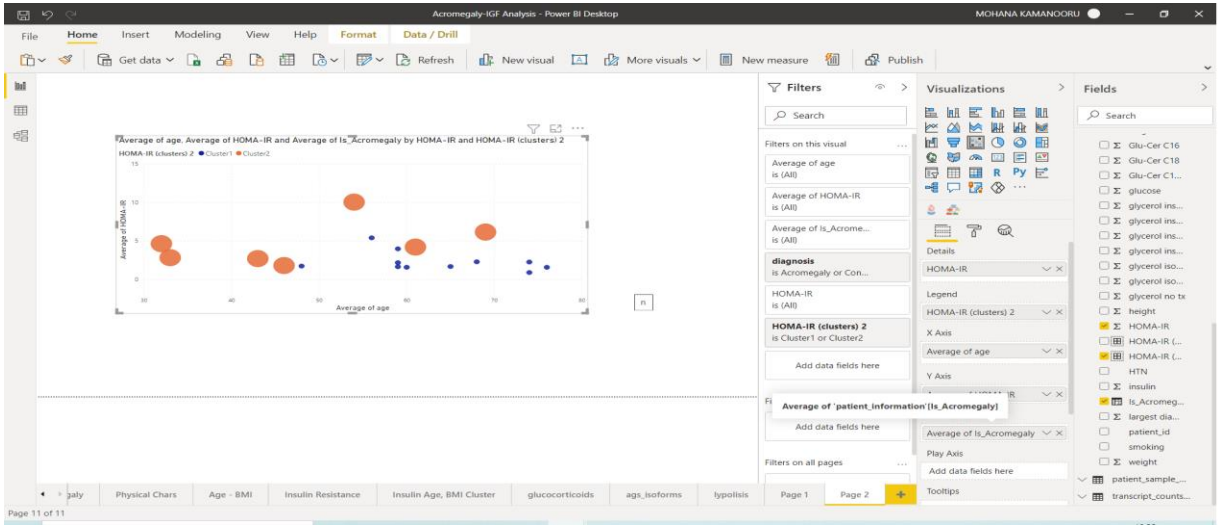


Now, drag and drop HOMA_IR in the Details tab under Visualization Pane.

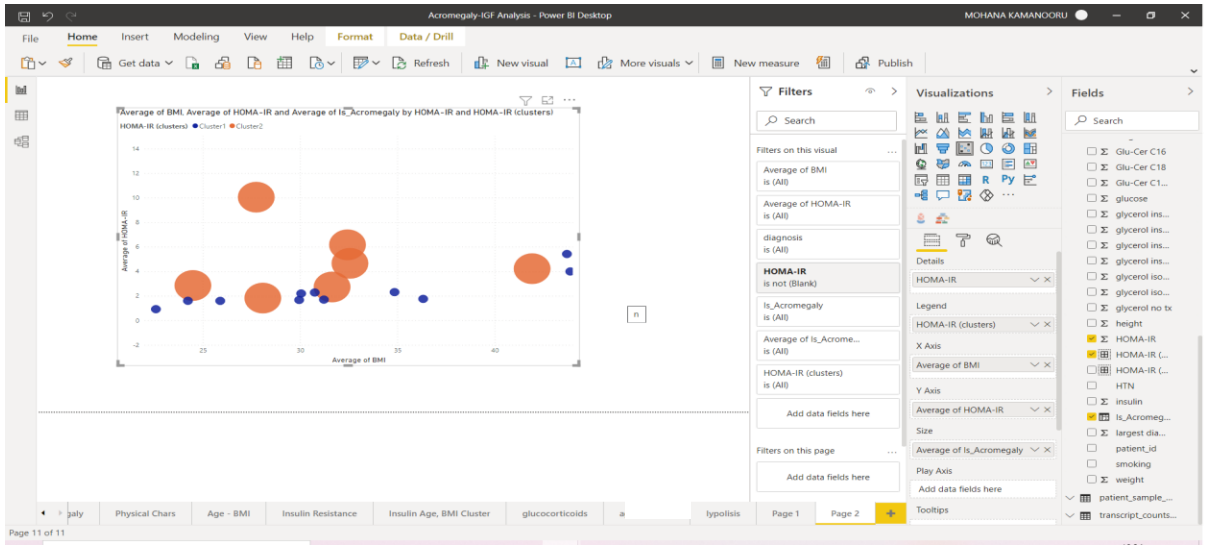


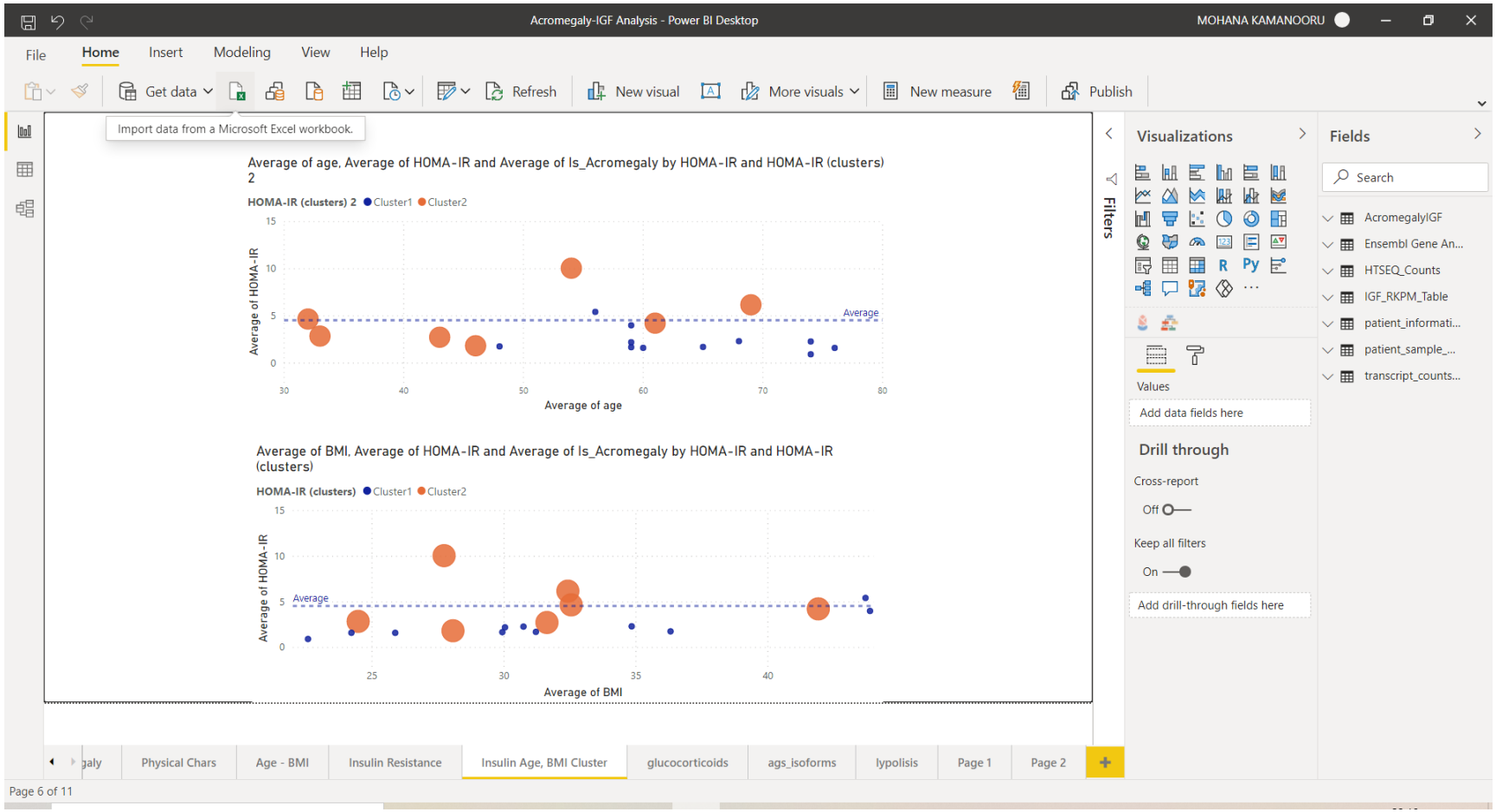


Drag and drop the newly created column **is_Acromegaly** in size and select Average.



Follow the same process to plot the graph with BMI from the patient_information Table.





5.1. Analytics

From both the scatterplots, the clustering algorithm seems to be more appropriate. The algorithm sectioned two clusters, where cluster 2 resembles the Acromegaly patient group and cluster1 represents the Control group.

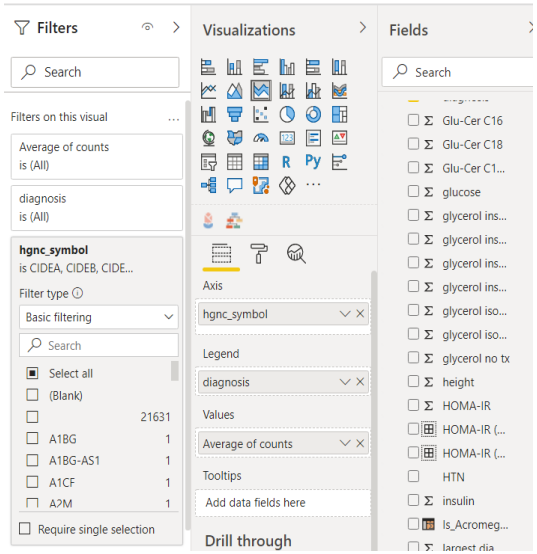
5.2. Findings

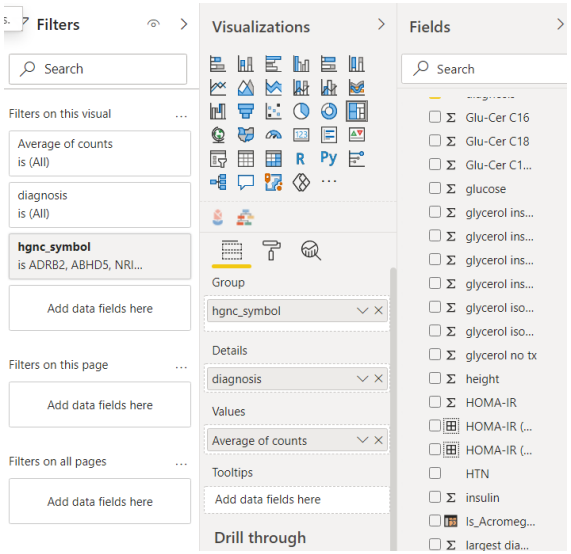
Acromegaly patients' data has different characteristics and Control patients' data has different characteristics. These are more accurately predicted by the machine learning algorithm for the current dataset sample.

6. Lipolysis, Glucocorticoids, and Isoforms Analysis

6.1. Lipolysis

Also, plot the Treemap and Stacked area chart to examine the Lipolysis Genes and Regulators among the patient groups. For Lipolysis Regulators select 'CIDEA', 'CIDEB', 'CIDEC', and 'G0S2' from Filters Pane for hgnc_symbol under basic filtering. For Lipolysis Genes select "ABHD5", "NRIP1", "ADRB1", "ADRB2", and "ADRB3".





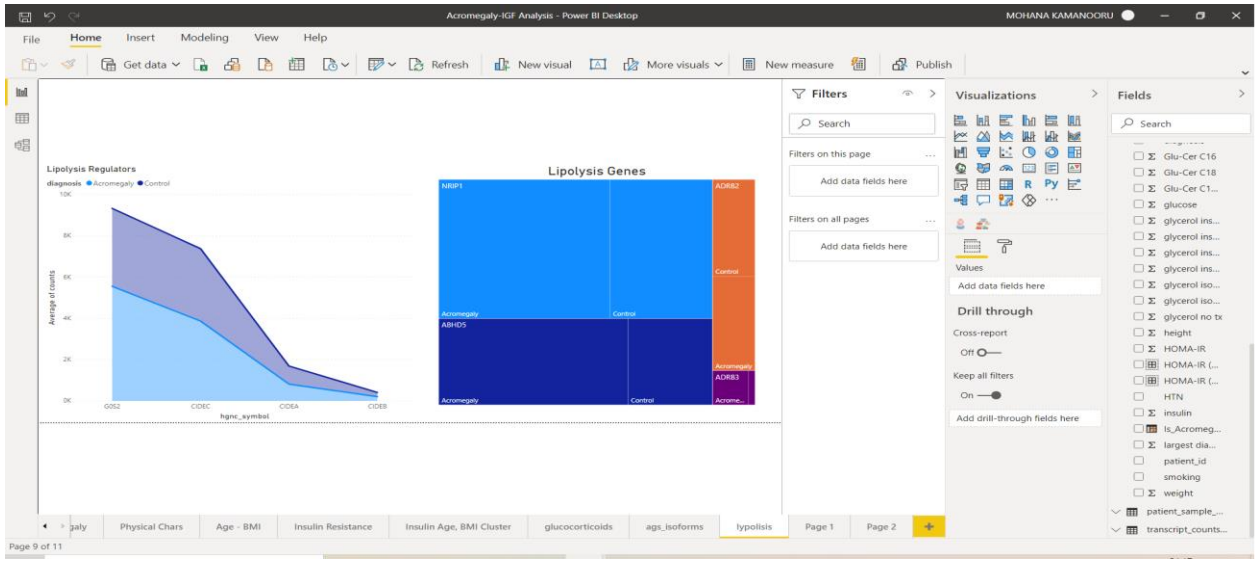


Figure 7 Lipolysis Plots

6.2. Glucocorticoids

Create a horizontal plot for an average of “counts” from HTSEQ_Counts table by “diagnosis” column from patient_information Table, and filter by hgnc_symbol column with values (“HSD11B1", "HSD11B2", "NR3C1", "NR3C2") to evaluate glucocorticoids relativity if any between patient groups.

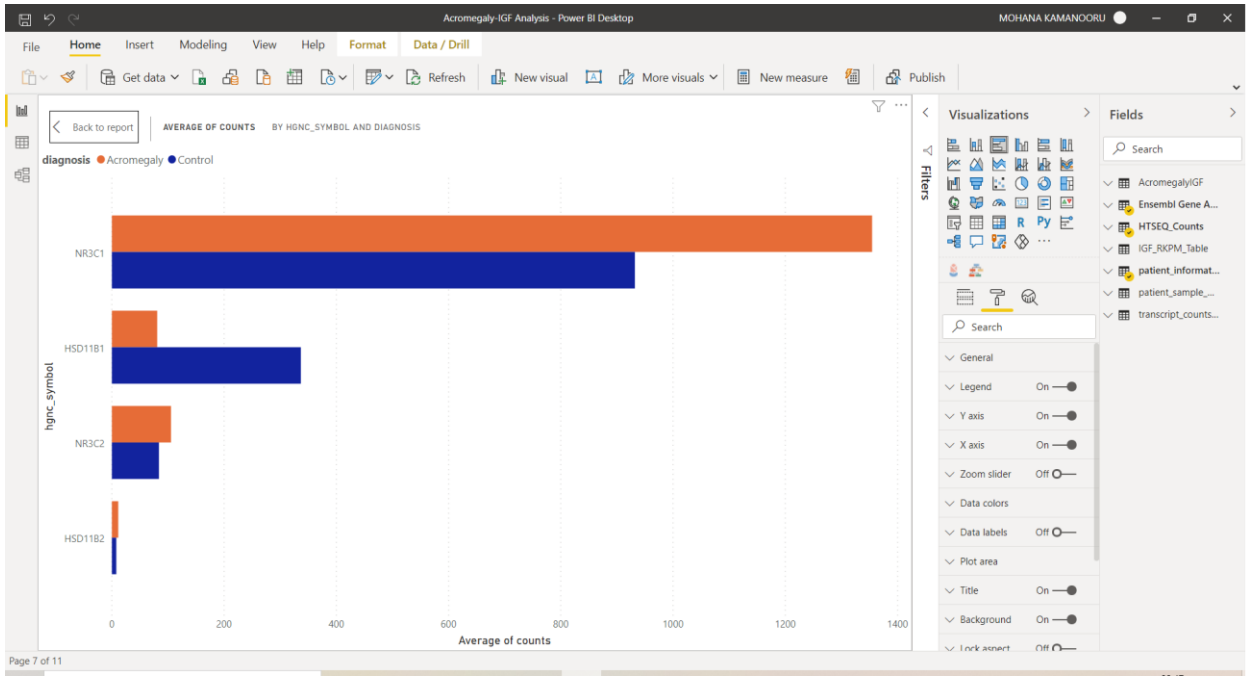


Figure 8 Glucocorticoids Analysis Plot

6.3. Isoforms

Similarly, plot the vertical bar graph for an average of “counts” from HTSEQ_Counts table by “diagnosis” column from patient_information Table, and filter by hgnc_symbol column with values (“GPSM2", "GPSM1", "GPSM3") to evaluate AGS protein isoforms in both patient groups.

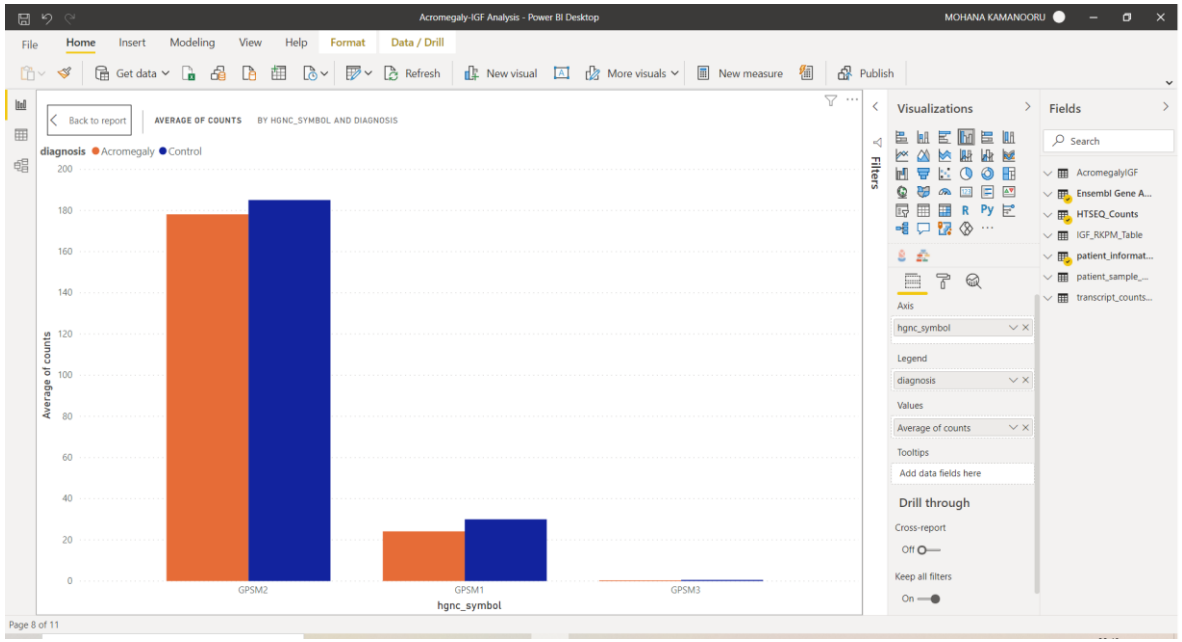


Figure 9 Isoforms Analysis plot

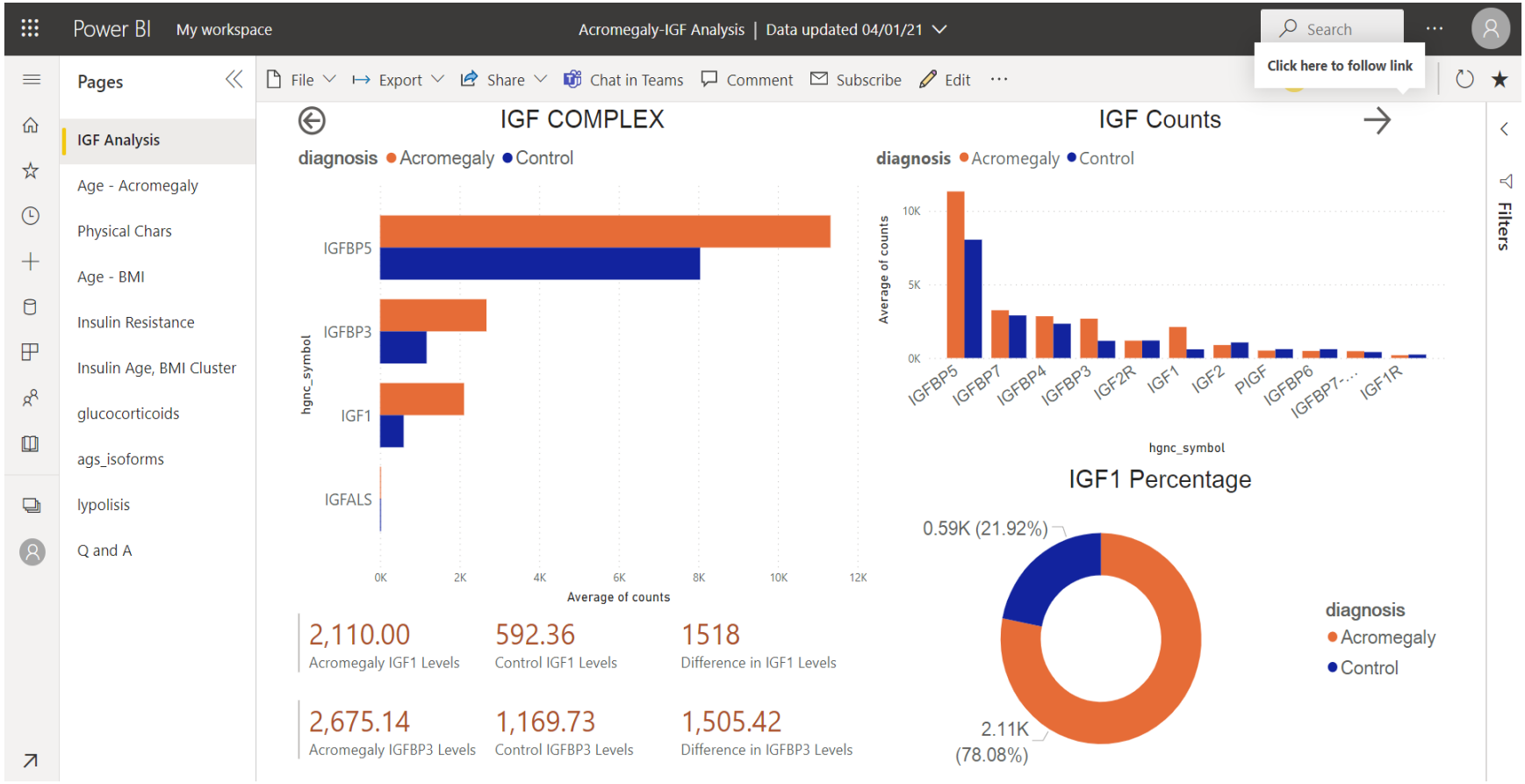
6.1. Analysis

From the plots above, an increase in gene counts has been observed for acromegaly. Changes in the Lipolysis gene and regulators are related to insulin absorption levels. Glucocorticoids increase affect the glucose levels and decrease in AGS protein isoform in patient groups.

6.2. Findings

In this research, it is observed that the patients with Acromegaly have high glucose and insulin resistance, also have low insulin sensitivity. This also supports the high HOMA-IR score in patients and the observations in the insulin analysis section.

Power BI Report View



Dashboard View

Conclusions

Analyzing Acromegaly and IGF from the dataset captured from opensource. The average IGF1 levels are 3.5 times higher in Acromegaly patients. IGF1 percentage is 78% for Acromegaly and 22% for control patients. Also, Acromegaly patients have higher insulin and higher lipolysis. Acromegaly patients tend to show higher averages for BMI and height. The probability of Acromegaly might increase with age. Since the analyzed data is a smaller sample, this may not be true for more significant models and real-time scenarios. Weight and Abdominal circumferences do not tend to change in considerable amounts according to the research dataset.

References

[1] "Acromegaly - NHS." https://www.nhs.uk/conditions/acromegaly/ (accessed Jan. 05, 2021).

[2] "Acromegaly - Symptoms, and causes - Mayo Clinic." https://www.mayoclinic.org/diseases-conditions/acromegaly/symptoms-causes/syc-20351222 (accessed Jan. 05, 2021).

[3] J. O. L. Jørgensen *et al.*, "GH receptor signaling in skeletal muscle and adipose tissue in human subjects following exposure to an intravenous GH bolus," *American Journal of Physiology - Endocrinology and Metabolism*, vol. 291, no. 5, 2006, doi: 10.1152/ajpendo.00024.2006.

[4] J. S. Huo *et al.*, "Profiles of growth hormone (GH)-regulated genes reveal time-dependent responses and identify a mechanism for regulation of activating transcription factor 3 by GH," *Journal of Biological Chemistry*, vol. 281, no. 7, pp. 4132–4141, Feb. 2006, doi: 10.1074/jbc.M508492200.

[5] J. Bolinder, J. Ostman, S. Werner, and P. Arner, "Insulin action in human adipose tissue in acromegaly," *Journal of Clinical Investigation*, vol. 77, no. 4, pp. 1201–1206, 1986, doi: 10.1172/JCI112422.

[6] "Gene Expression Signature in Adipose Tissue of Acromegaly Patients." https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0129359 (accessed Jan. 05, 2021).

[7] J. Ayuk and M. C. Sheppard, “Growth hormone and its disorders,” *Postgraduate Medical Journal*, vol. 82, no. 963. pp. 24–30, Jan. 2006, doi: 10.1136/pgmj.2005.036087.

[8] I. M. Holdaway and C. Rajasoorya, “Epidemiology of Acromegaly,” *Pituitary*, vol. 2, no. 1, pp. 29–41, 1999, doi: 10.1023/A:1009965803750.

[9] K. Tanimoto, N. Hizuka, I. Fukuda, K. Takano, and T. Hanafusa, “The influence of age on the GH-IGF1 axis in patients with acromegaly,” *European Journal of Endocrinology*, vol. 159, no. 4, pp. 375–379, Oct. 2008, doi: 10.1530/EJE-08-0243.

[10] I. Hochberg, Q. T. Tran, A. L. Barkan, A. R. Saltiel, W. F. Chandler, and D. Bridges, “Gene Expression Signature in Adipose Tissue of Acromegaly Patients,” *PLOS ONE*, vol. 10, no. 6, p. e0129359, Jun. 2015, doi: 10.1371/journal.pone.0129359.

[11] A. Vijayakumar, R. Novosyadlyy, Y. J. Wu, S. Yakar, and D. LeRoith, “Biological effects of growth hormone on carbohydrate and lipid metabolism,” *Growth Hormone and IGF Research*, vol. 20, no. 1. pp. 1–7, Feb. 2010, doi: 10.1016/j.ghir.2009.09.002.

[12] R. B. Simsolo, S. Ezzat, J. M. Ong, M. Saghizadeh, and P. A. Kern, “Effects of acromegaly treatment and growth hormone on adipose tissue lipoprotein lipase,” *Journal of Clinical Endocrinology and Metabolism*, vol. 80, no. 11, pp. 3233–3238, 1995, doi: 10.1210/jcem.80.11.7593431.

[13] S. Anders, P. T. Pyl, and W. Huber, “HTSeq-A Python framework to work with high-throughput sequencing data,” *Bioinformatics*, vol. 31, no. 2, pp. 166–169, Jan. 2015, doi: 10.1093/bioinformatics/btu638.