DATA CLEANING

DR DANNY POO

BIG DATA ANALYTICS AND VISUALISATION

Data Cleaning Tasks

Change data format

- Text
- Number

Date and Time

Remove duplicate data

Remove rows with missing data

Remove conflicting data

Find and replace texts

Bin data into buckets

Find outliers

Overview of Data File

Data file: dc-appointment.xlsx

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	PatientId	AppointmentID	Gender	ScheduleDay	AppointmentDay	Age	Neighbourhood	FamilyAllowance	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	No-show
2	2.98725E+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
3	5.58998E+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
4	2.97836E+12	5671456	F	2016-05-06T17:10:31Z	2016-05-12T00:00:00Z	21	JARDIM DA PENHA	0	0	0	0	0	1	No
5	6.32644E+12	5662931	F	2016-05-05T09:14:10Z	2016-05-05T00:00:00Z	55	JARDIM DA PENHA	0	1	0	0	0	0	No
6	5.55834E+13	5608066	M	2016-04-20T13:12:17Z	2016-05-12T00:00:00Z		JARDIM DA PENHA	0	1	0	0	0	0	No
7	8.48511E+14	5562224	M	2016-04-08T12:56:06Z	2016-05-06T00:00:00Z	77	JARDIM DA PENHA	0	1	0	1	0	1	No
8	5.79444E+12	5587385	F	2016-04-15T08:37:18Z	2016-05-13T00:00:00Z	57	JARDIM DA PENHA	0	0	0	0	0	0	No
9	2.1216E+13	5617350	F	2016-04-25T15:58:50Z	2016-05-20T00:00:00Z	76	JARDIM DA PENHA	0	1	1	0	0	0	No
10	1.67235E+13	5668727	F	2016-05-06T09:38:36Z	2016-05-06T00:00:00Z	60	JARDIM DA PENHA	0	0	0	0	0	0	No
11	9.95197E+14	5695235	F	2016-05-13T09:47:09Z	2016-05-13T00:00:00Z	38	JARDIM DA PENHA	0	0	0	0	0	0	No
12	6.73686E+12	5668957	F	2016-05-06T10:02:59Z	2016-05-06T00:00:00Z	60	JARDIM DA PENHA	0	1	0	0	0	0	No
13	6.26614E+13	5695655	F	2016-05-13T10:41:09Z	2016-05-13T00:00:00Z	54	JARDIM DA PENHA	0	1	0	0	0	0	No
14	9.94496E+11	5725320	M	2016-05-20T10:54:26Z	2016-05-20T00:00:00Z	20	JARDIM DA PENHA	0	0	0	0	0	0	No
15	8.33917E+13	5563722	F	2016-04-08T18:18:01Z	2016-05-06T00:00:00Z	85	JARDIM DA PENHA	0	1	0	0	0	0	No

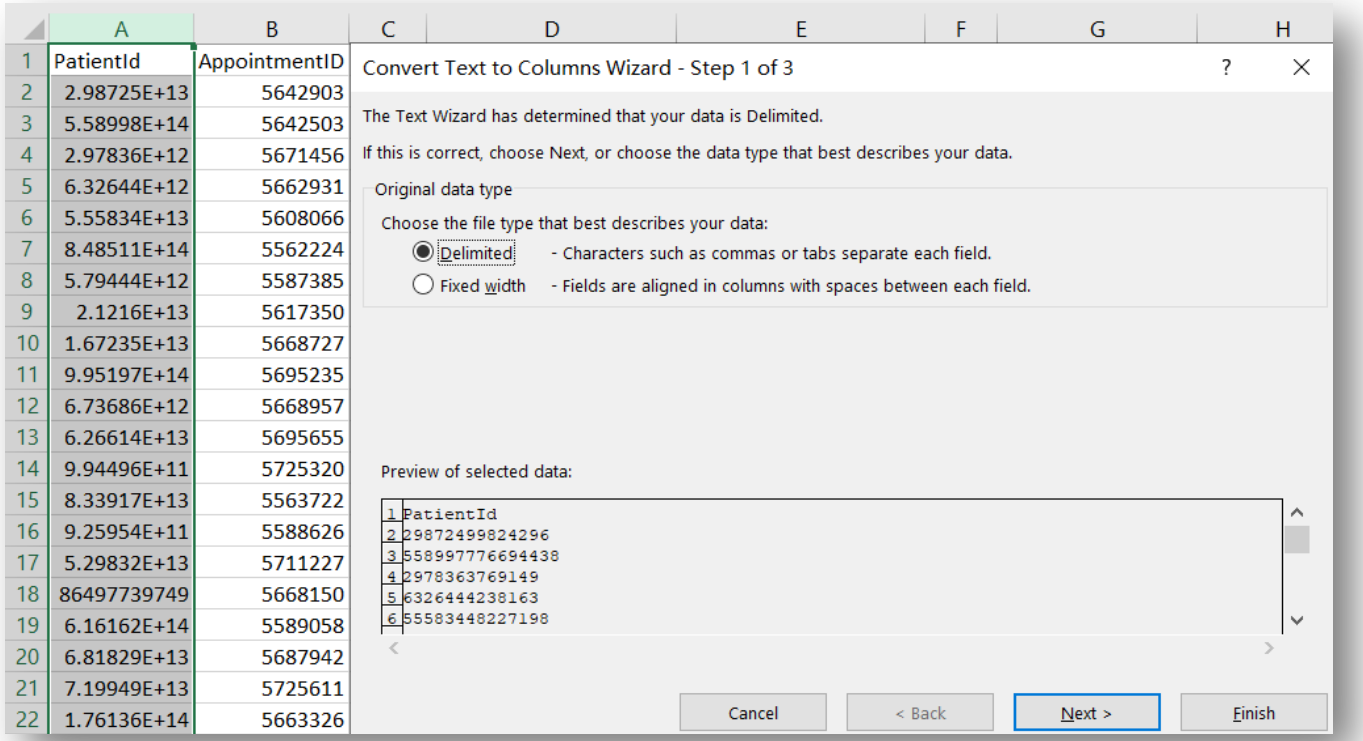
6569	5.88313E+14	5763362	F	2016-06-02T08:53:13Z	2016-06-06T00:00:00Z	60	JARDIM CAMBURI	0	0	0	0	0	0	Yes
6570	1.56131E+13	5779070	M	2016-06-06T17:42:01Z	2016-06-06T00:00:00Z	62	JARDIM CAMBURI	0	0	0	0	0	0	No
6571	19314223565	5767006	F	2016-06-02T16:13:58Z	2016-06-07T00:00:00Z	55	JARDIM CAMBURI	0	0	0	0	0	1	No
6572	5.96576E+12	5771390	F	2016-06-03T12:45:35Z	2016-06-07T00:00:00Z	56	JARDIM CAMBURI	0	0	0	0	0	1	No
6573	7.82544E+12	5767544	F	2016-06-02T18:32:06Z	2016-06-07T00:00:00Z	60	JARDIM CAMBURI	0	0	0	0	0	1	No
6574	5.98675E+14	5770793	M	2016-06-03T10:57:50Z	2016-06-07T00:00:00Z	44	JARDIM CAMBURI	0	0	0	0	0	1	No

Initial insights:
6573 rows, 14 columns

CHANGE DATA FORMAT

Scientific Data Format for PatientId Unsuitable: Change to Text Format ..1/3

	A
1	PatientId
2	2.98725E+13
3	5.58998E+14
4	2.97836E+12
5	6.32644E+12
6	5.55834E+13
7	8.48511E+14
8	5.79444E+12
9	2.1216E+13
10	1.67235E+13
11	9.95197E+14
12	6.73686E+12
13	6.26614E+13
14	9.94496E+11
15	8.33917E+13
16	9.25954E+11
17	5.29832E+13
18	86497739749
19	6.16162E+14
20	6.81829E+13



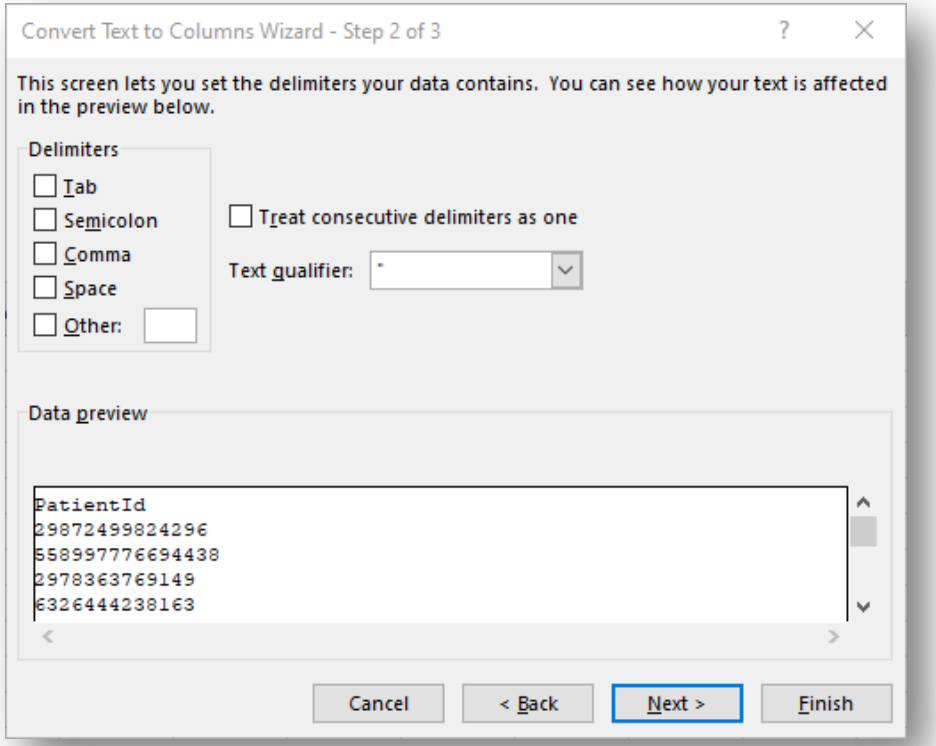
	A	B
1	PatientId	AppointmentID
2	2.98725E+13	5642903
3	5.58998E+14	5642503
4	2.97836E+12	5671456
5	6.32644E+12	5662931
6	5.55834E+13	5608066
7	8.48511E+14	5562224
8	5.79444E+12	5587385
9	2.1216E+13	5617350
10	1.67235E+13	5668727
11	9.95197E+14	5695235
12	6.73686E+12	5668957
13	6.26614E+13	5695655
14	9.94496E+11	5725320
15	8.33917E+13	5563722
16	9.25954E+11	5588626
17	5.29832E+13	5711227
18	86497739749	5668150
19	6.16162E+14	5589058
20	6.81829E+13	5687942
21	7.19949E+13	5725611
22	1.76136E+14	5663326

How?

Click Data tab > Text to Columns > Delimited > Next

Scientific Data Format for PatientId Unsuitable: Change to Text Format ..2/3

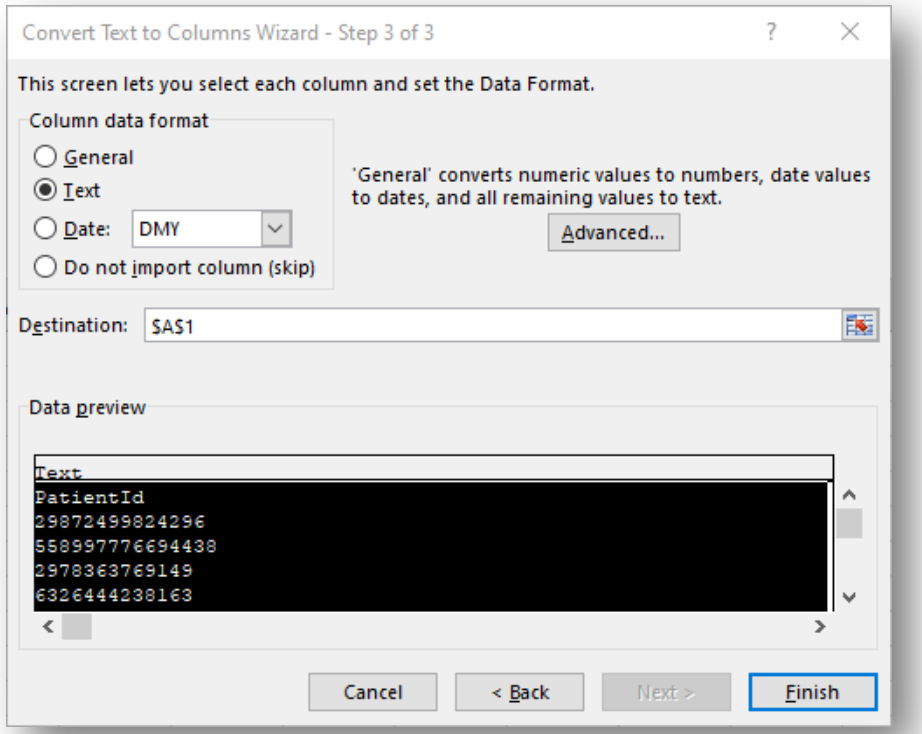
	A
1	PatientId
2	2.98725E+13
3	5.58998E+14
4	2.97836E+12
5	6.32644E+12
6	5.55834E+13
7	8.48511E+14
8	5.79444E+12
9	2.1216E+13
10	1.67235E+13
11	9.95197E+14
12	6.73686E+12
13	6.26614E+13
14	9.94496E+11
15	8.33917E+13
16	9.25954E+11
17	5.29832E+13
18	86497739749
19	6.16162E+14
20	6.81829E+13



How?
Click Next

Scientific Data Format for PatientId Unsuitable: Change to Text Format ..3/3

	A
1	PatientId
2	2.98725E+13
3	5.58998E+14
4	2.97836E+12
5	6.32644E+12
6	5.55834E+13
7	8.48511E+14
8	5.79444E+12
9	2.1216E+13
10	1.67235E+13
11	9.95197E+14
12	6.73686E+12
13	6.26614E+13
14	9.94496E+11
15	8.33917E+13
16	9.25954E+11
17	5.29832E+13
18	86497739749
19	6.16162E+14
20	6.81829E+13

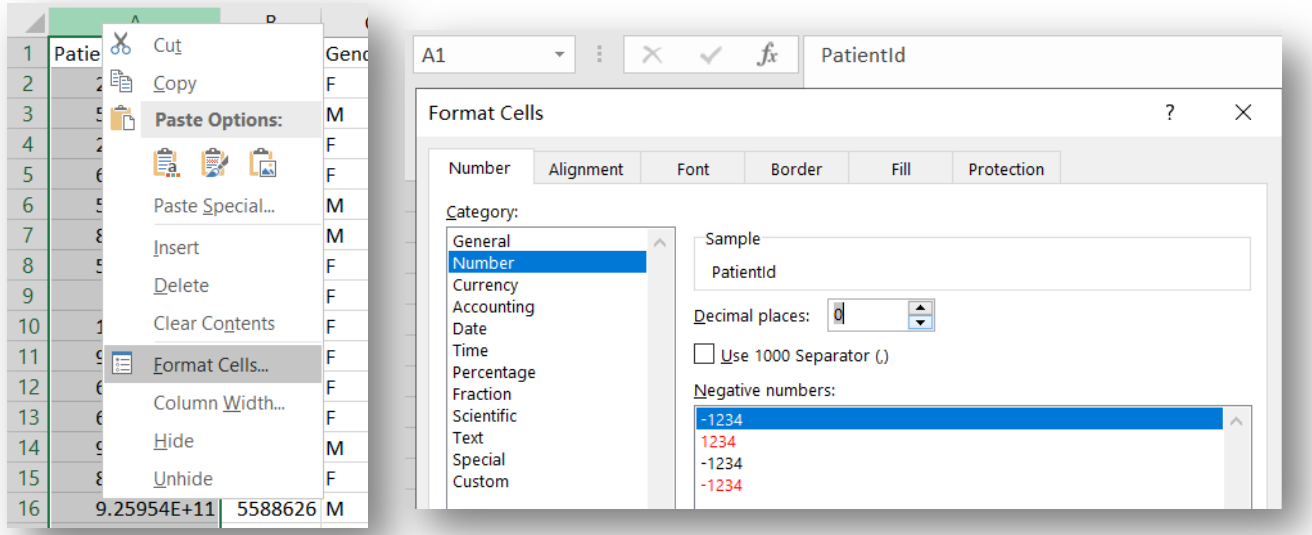


How?
Select Text > Finish

	A	B	C
1	PatientId	AppointmentID	Gender
2	29872499824296	5642903	F
3	558997776694438	5642503	M
4	2978363769149	5671456	F
5	6326444238163	5662931	F
6	55583448227198	5608066	M
7	848511136446923	5562224	M
8	5794444469345	5587385	F
9	21215964327482	5617350	F
10	16723534136818	5668727	F
11	995196886477797	5695235	F
12	6736855292851	5668957	F
13	62661375954544	5695655	F
14	994495686222	5725320	M
15	83391743889612	5563722	F
16	925953619658	5588626	M
17	52983244522738	5711227	F
18	86497739749	5668150	F
19	616161776271474	5589058	F
20	68182861542773	5687942	F
21	71994897663243	5725611	F
22	176136196269528	5663326	F
23	68838234725772	5664257	M

Scientific Data Format for PatientId Unsuitable: Change to Number Format

	A
1	PatientId
2	2.98725E+13
3	5.58998E+14
4	2.97836E+12
5	6.32644E+12
6	5.55834E+13
7	8.48511E+14
8	5.79444E+12
9	2.1216E+13
10	1.67235E+13
11	9.95197E+14
12	6.73686E+12
13	6.26614E+13
14	9.94496E+11
15	8.33917E+13
16	9.25954E+11
17	5.29832E+13
18	86497739749
19	6.16162E+14
20	6.81829E+13



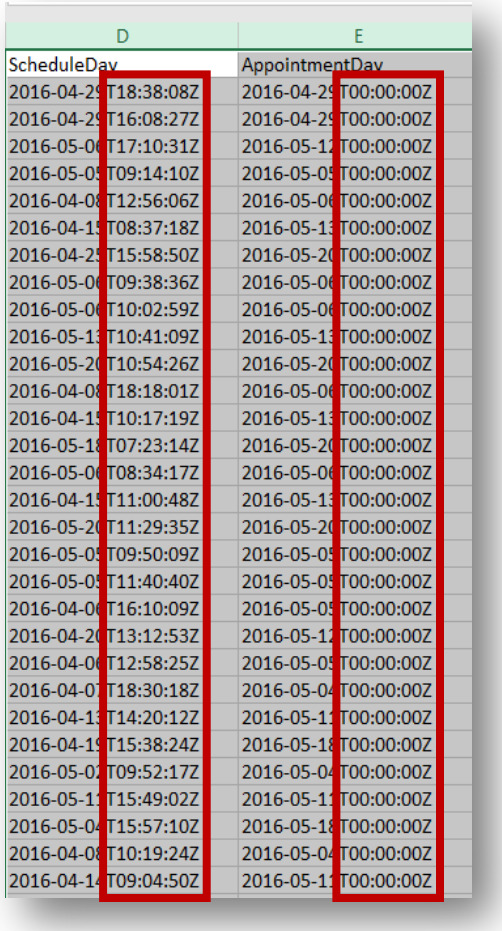
How?

Select PatientId column > Right click on column > Format Cells > Number > Decimal places:0

	A
1	PatientId
2	2987249824296
3	558997776694438
4	2978363769149
5	6326444238163
6	55583448227198
7	848511136446923
8	5794444469345
9	21215964327482
10	16723534136818
11	995196886477797
12	6736855292851
13	62661375954544
14	994495686222
15	83391743889612
16	925953619658
17	52983244522738
18	86497739749
19	616161776271474
20	68182861542773

DATE AND TIME

ScheduleDay and AppointmentDay



D	E
ScheduleDay	AppointmentDay
2016-04-29T18:38:08Z	2016-04-29T00:00:00Z
2016-04-29T16:08:27Z	2016-04-29T00:00:00Z
2016-05-06T17:10:31Z	2016-05-12T00:00:00Z
2016-05-05T09:14:10Z	2016-05-05T00:00:00Z
2016-04-08T12:56:06Z	2016-05-06T00:00:00Z
2016-04-15T08:37:18Z	2016-05-13T00:00:00Z
2016-04-29T15:58:50Z	2016-05-20T00:00:00Z
2016-05-06T09:38:36Z	2016-05-06T00:00:00Z
2016-05-06T10:02:59Z	2016-05-06T00:00:00Z
2016-05-13T10:41:09Z	2016-05-13T00:00:00Z
2016-05-20T10:54:26Z	2016-05-20T00:00:00Z
2016-04-08T18:18:01Z	2016-05-06T00:00:00Z
2016-04-15T10:17:19Z	2016-05-13T00:00:00Z
2016-05-18T07:23:14Z	2016-05-20T00:00:00Z
2016-05-06T08:34:17Z	2016-05-06T00:00:00Z
2016-04-15T11:00:48Z	2016-05-13T00:00:00Z
2016-05-20T11:29:35Z	2016-05-20T00:00:00Z
2016-05-05T09:50:09Z	2016-05-05T00:00:00Z
2016-05-05T11:40:40Z	2016-05-05T00:00:00Z
2016-04-06T16:10:09Z	2016-05-05T00:00:00Z
2016-04-20T13:12:53Z	2016-05-12T00:00:00Z
2016-04-06T12:58:25Z	2016-05-05T00:00:00Z
2016-04-07T18:30:18Z	2016-05-04T00:00:00Z
2016-04-13T14:20:12Z	2016-05-11T00:00:00Z
2016-04-15T15:38:24Z	2016-05-18T00:00:00Z
2016-05-02T09:52:17Z	2016-05-04T00:00:00Z
2016-05-11T15:49:02Z	2016-05-11T00:00:00Z
2016-05-04T15:57:10Z	2016-05-18T00:00:00Z
2016-04-08T10:19:24Z	2016-05-04T00:00:00Z
2016-04-14T09:04:50Z	2016-05-11T00:00:00Z

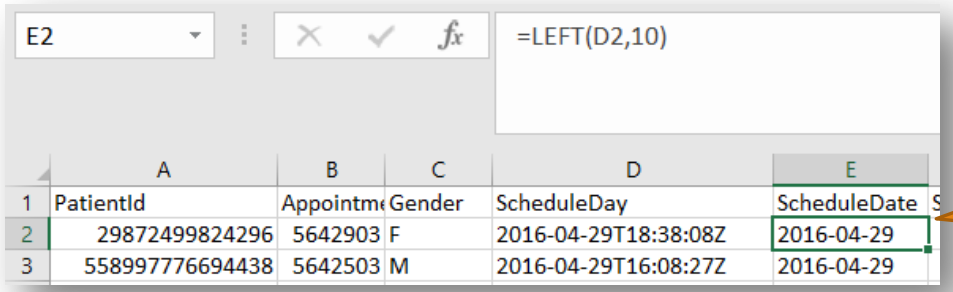
Split ScheduleDay into ScheduleDate and ScheduleTime
Remove Time part of AppointmentDay

Add New Columns: ScheduleDate, ScheduleTime and AppointmentDate

Add ScheduleDate, ScheduleTime and AppointmentDate columns

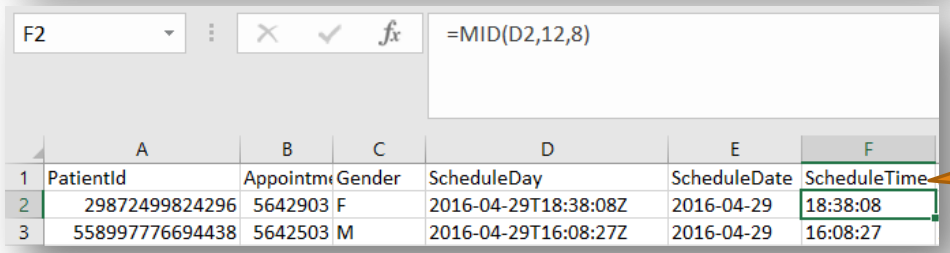
	A	B	C	D	E	F	G	H	I
1	PatientId	AppointmentID	Gender	ScheduleDay	ScheduleDate	ScheduleTime	AppointmentDay	AppointmentDate	Age
2	29872499824296	5642903	F	2016-04-29T18:38:08Z			2016-04-29T00:00:00Z		62
3	558997776694438	5642503	M	2016-04-29T16:08:27Z			2016-04-29T00:00:00Z		56
4	2978363769149	5671456	F	2016-05-06T17:10:31Z			2016-05-12T00:00:00Z		21
5	6326444238163	5662931	F	2016-05-05T09:14:10Z			2016-05-05T00:00:00Z		55
6	55583448227198	5608066	M	2016-04-20T13:12:17Z			2016-05-12T00:00:00Z		
7	848511136446923	5562224	M	2016-04-08T12:56:06Z			2016-05-06T00:00:00Z		77

Split ScheduleDay and AppointmentDay

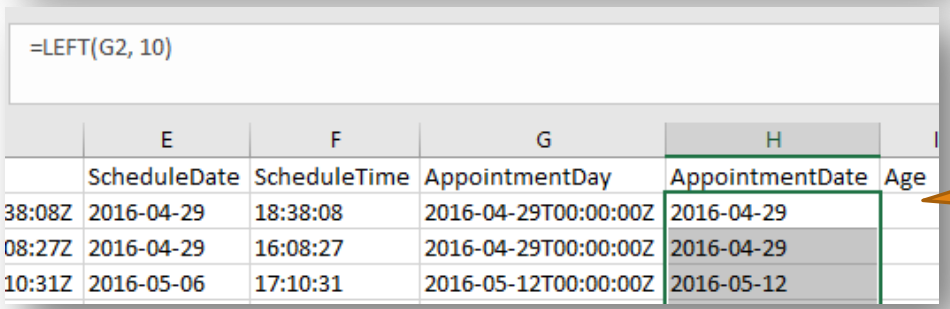


Use Split functions: LEFT, MID

ScheduleDate
=LEFT(ScheduleDay, 10)

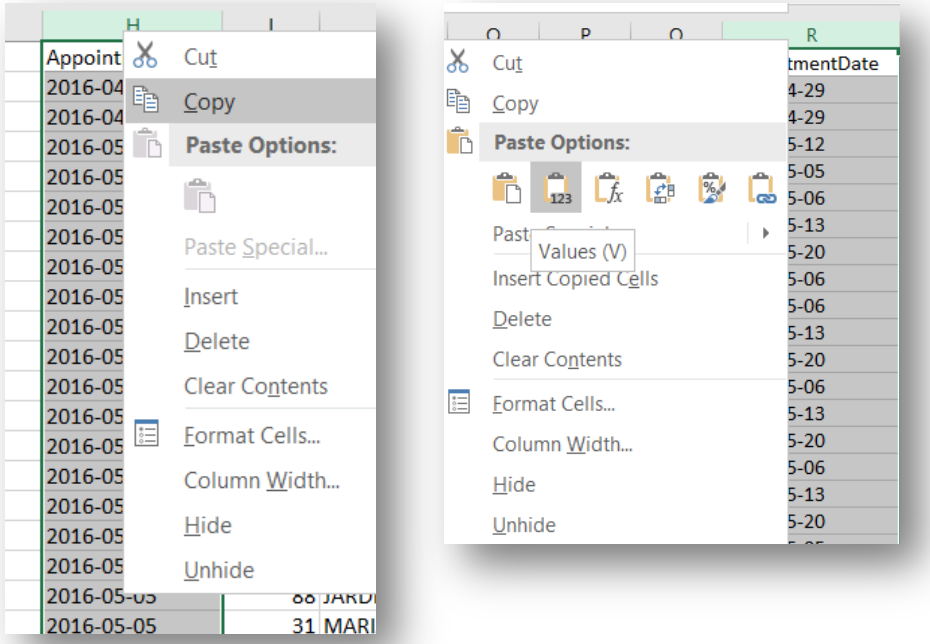


ScheduleTime
=MID(ScheduleDay, 12, 8)



AppointmentDate
=LEFT(AppointmentDay, 10)

Copy Value AppointmentDate, ScheduleDate and ScheduleTime

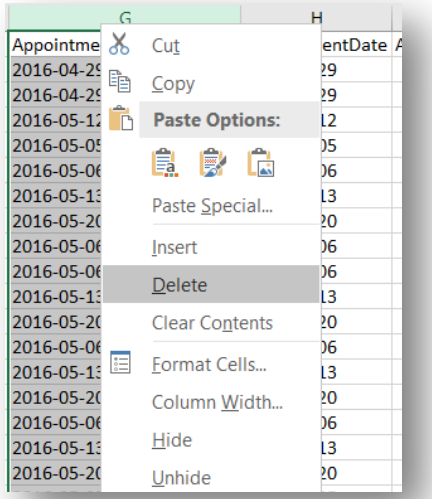


How?
Select AppointmentDate column > Right click Copy > goto column R > Paste value

How?
Select ScheduleDate column > Right click Copy > goto column S > Paste value

How?
Select ScheduleTime column > Right click Copy > goto column T > Paste value

Delete Old Columns



How?

Select ScheduleDay, ScheduleDate, ScheduleTime, AppointmentDay, AppointmentDate columns > Right click Delete

Resulting columns for AppointmentDate, ScheduleDate and ScheduleTime

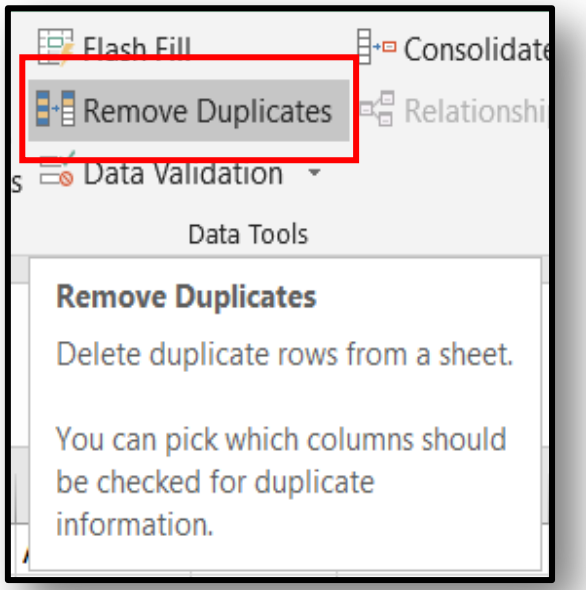
M	N	O
AppointmentDate	ScheduleDate	ScheduleTime
2016-04-29	2016-04-29	18:38:08
2016-04-29	2016-04-29	16:08:27
2016-05-12	2016-05-06	17:10:31
2016-05-05	2016-05-05	09:14:10
2016-05-12	2016-04-20	13:12:17
2016-05-06	2016-04-08	12:56:06
2016-05-13	2016-04-15	08:37:18
2016-05-20	2016-04-25	15:58:50
2016-05-06	2016-05-06	09:38:36
2016-05-13	2016-05-13	09:47:09
2016-05-06	2016-05-06	10:02:59
2016-05-13	2016-05-13	10:41:09
2016-05-20	2016-05-20	10:54:26
2016-05-06	2016-04-08	18:18:01
2016-05-13	2016-04-15	10:17:19

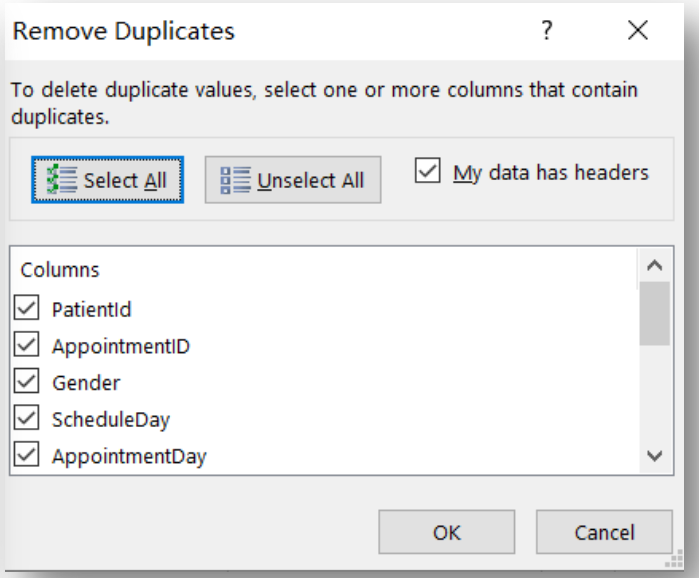
So Far,

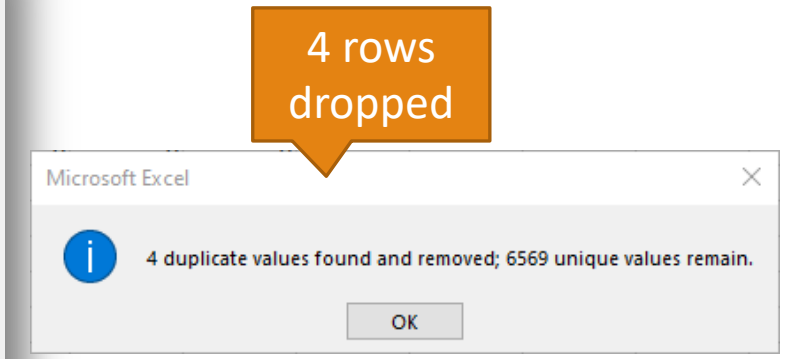
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	PatientId	AppointmentID	Gender	Age	Neighbourhood	FamilyAllowance	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	No-show	AppointmentDate	ScheduleDate	ScheduleTime
2	29872499824296	5642903	F	62	JARDIM DA PENHA	0	1	0	0	0	0	No	2016-04-29	2016-04-29	18:38:08
3	558997776694438	5642503	M	56	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-04-29	2016-04-29	16:08:27
4	2978363769149	5671456	F	21	JARDIM DA PENHA	0	0	0	0	0	1	No	2016-05-12	2016-05-06	17:10:31
5	6326444238163	5662931	F	55	JARDIM DA PENHA	0	1	0	0	0	0	No	2016-05-05	2016-05-05	09:14:10
6	55583448227198	5608066	M		JARDIM DA PENHA	0	1	0	0	0	0	No	2016-05-12	2016-04-20	13:12:17
7	848511136446923	5562224	M	77	JARDIM DA PENHA	0	1	0	1	0	1	No	2016-05-06	2016-04-08	12:56:06
8	5794444469345	5587385	F	57	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-13	2016-04-15	08:37:18
9	21215964327482	5617350	F	76	JARDIM DA PENHA	0	1	1	0	0	0	No	2016-05-20	2016-04-25	15:58:50
10	16723534136818	5668727	F	60	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-06	2016-05-06	09:38:36
11	995196886477797	5695235	F	38	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-13	2016-05-13	09:47:09
12	6736855292851	5668957	F	60	JARDIM DA PENHA	0	1	0	0	0	0	No	2016-05-06	2016-05-06	10:02:59
13	62661375954544	5695655	F	54	JARDIM DA PENHA	0	1	0	0	0	0	No	2016-05-13	2016-05-13	10:41:09
14	994495686222	5725320	M	20	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-20	2016-05-20	10:54:26
15	83391743889612	5563722	F	85	JARDIM DA PENHA	0	1	0	0	0	0	No	2016-05-06	2016-04-08	18:18:01
16	925953619658	5588626	M	62	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-13	2016-04-15	10:17:19
17	52983244522738	5711227	F	78	JARDIM DA PENHA	0		0	0	0	0	No	2016-05-20	2016-05-18	07:23:14
18	86497739749	5668150	F	28	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-06	2016-05-06	08:34:17
19	616161776271474	5589058	F	46	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-13	2016-04-15	11:00:48
20	68182861542773	5687942	F	77	JARDIM DA PENHA	0	1	0	0	0	0	Yes	2016-05-20	2016-05-11	17:27:44

REMOVE DUPLICATE DATA

Remove Duplicate Rows

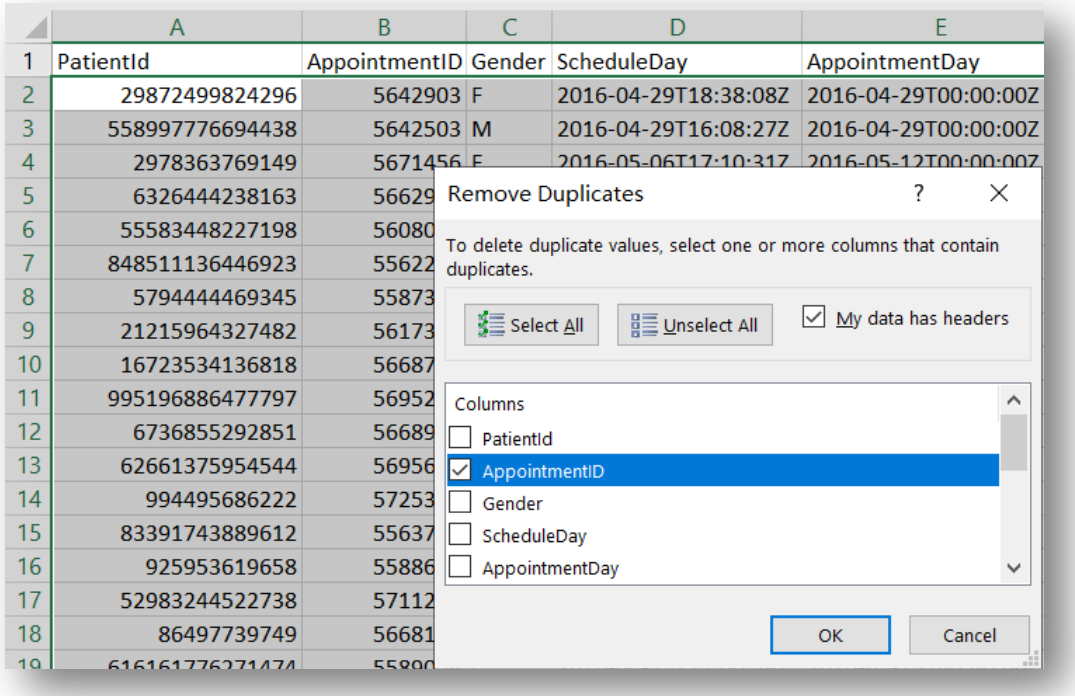


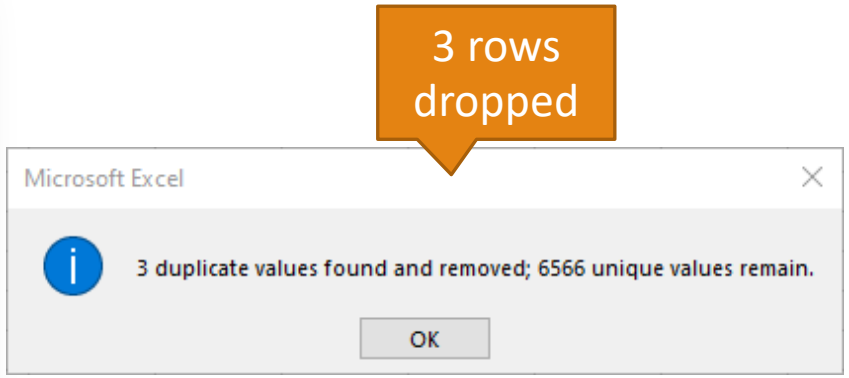




How?
Click any populated cell > Click Data tab > goto Data Tools > Remove Duplicates > Select All > OK > OK > save file

AppointmentID Cannot be Repeated



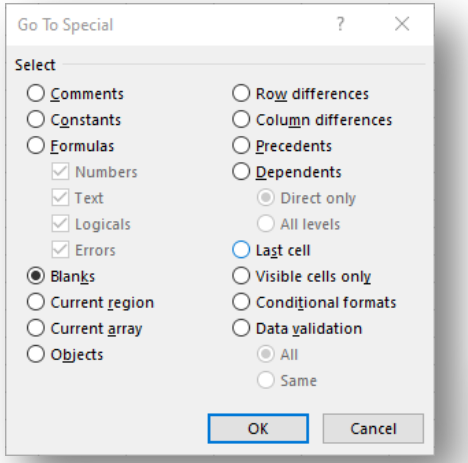


How?

Click any populated cell > Click Data tab > goto Data Tools > Remove Duplicates > Unselect All > check AppointmentID > OK > OK > save file

REMOVE ROWS WITH MISSING DATA

Identify Blanks



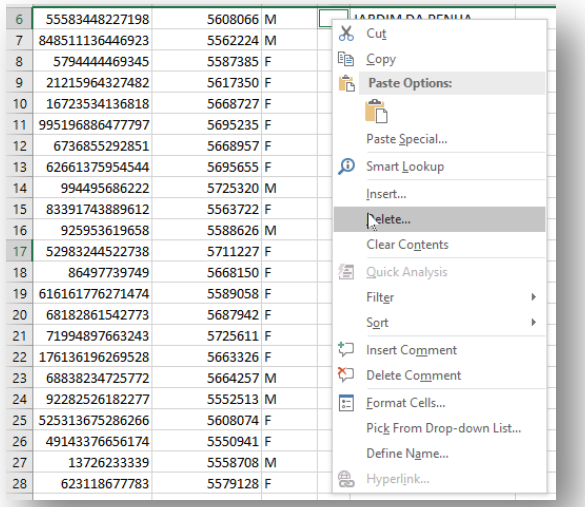
	A	B	C	D	E
1	PatientId	AppointmentID	Gender	Age	Neighbourhood
2	29872499824296	5642903	F	62	JARDIM DA PENHA
3	558997776694438	5642503	M	56	JARDIM DA PENHA
4	2978363769149	5671456	F	21	JARDIM DA PENHA
5	6326444238163	5662931	F	55	JARDIM DA PENHA
6	55583448227198	5608066	M		JARDIM DA PENHA
7	848511136446923	5562224	M	77	JARDIM DA PENHA

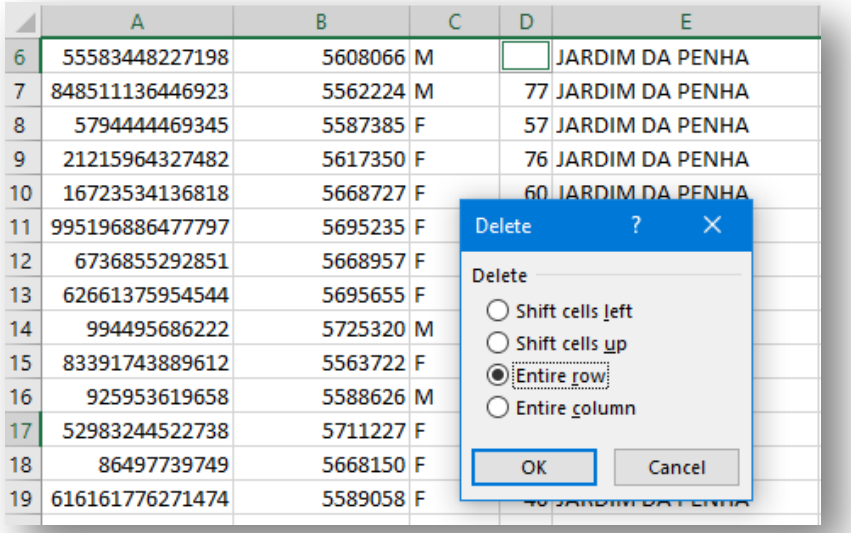
92	88276759865225	5622866	M	64	JARDIM DA PENHA	0	0	0	0
93	14218566861127	5647487	F	19	JARDIM DA PENHA	0	0	0	0
94	3983247493595	5675162	F		JARDIM DA PENHA	0	0	0	0
95	465468729494675	5746026	F	53	JARDIM DA PENHA	0	1	0	0
96	64477393396948	5644859	F	77	JARDIM DA PENHA	0	1		0
97	1559462794197	5569356	F	24	JARDIM DA PENHA	0	0	0	0
98	23133117288256	5623628	F	10	JARDIM DA PENHA	0	0	0	0
99	52188695613681	5647656	M	21	JUCUTUQUARA	0	0	0	0

How?

Click any populated area > Click Home tab > goto Editing tab > Click Find & Select > Go To Special > Blanks > OK

Remove Identified Blanks





	A	B	C	D	E
6	55583448227198	5608066	M		JARDIM DA PENHA
7	848511136446923	5562224	M	77	JARDIM DA PENHA
8	5794444469345	5587385	F	57	JARDIM DA PENHA
9	21215964327482	5617350	F	76	JARDIM DA PENHA
10	16723534136818	5668727	F	60	JARDIM DA PENHA
11	995196886477797	5695235	F		
12	6736855292851	5668957	F		
13	62661375954544	5695655	F		
14	994495686222	5725320	M		
15	83391743889612	5563722	F		
16	925953619658	5588626	M		
17	52983244522738	5711227	F		
18	86497739749	5668150	F		
19	616161776271474	5589058	F		

6550	5965757389437	5771390	F	56	JARDIM CAMBURI
6551	7825443833792	5767544	F	60	JARDIM CAMBURI
6552	598674846832672	5770793	M	44	JARDIM CAMBURI
6553					

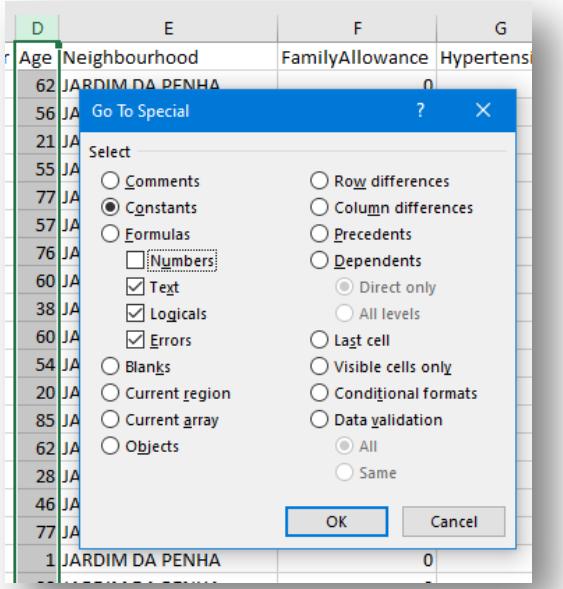
Last row number is 6552

How?

Click on any of the identified blank > Right click Delete > Entire row > OK

Find Other Missing Data

Non-numeric data in Age column can be considered as missing data

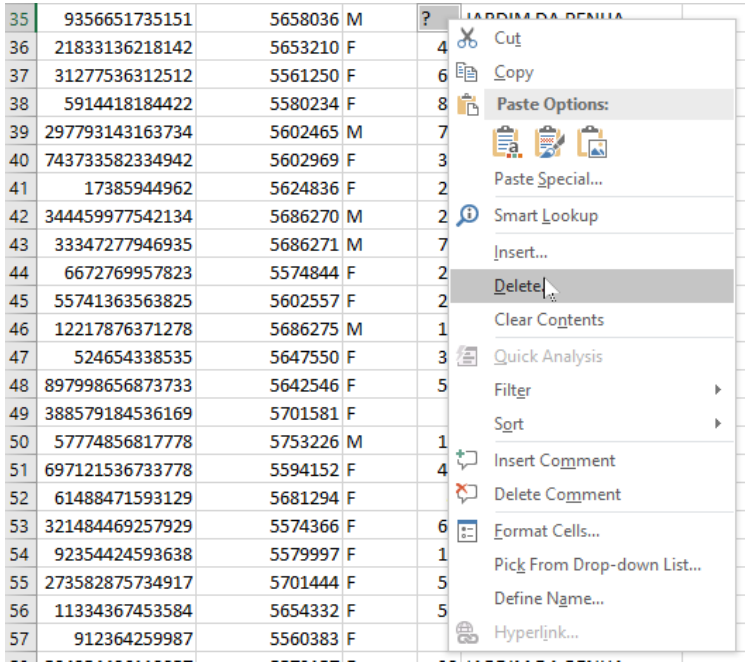


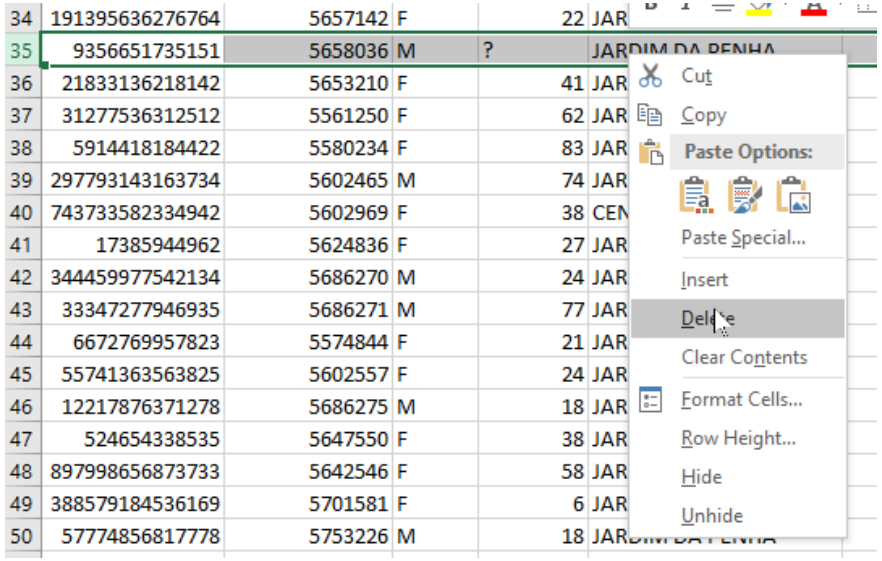
30	49821955232	5660262	F	51	JARDIM DA PENHA
31	3859455169236	5561244	M	66	JARDIM DA PENHA
32	75921919233878	5582208	F	76	JARDIM DA PENHA
33	929297341217427	5602736	M	59	JARDIM DA PENHA
34	191395636276764	5657142	F	22	JARDIM DA PENHA
35	9356651735151	5658036	M	?	JARDIM DA PENHA
36	21833136218142	5653210	F	41	JARDIM DA PENHA
37	31277536312512	5561250	F	62	JARDIM DA PENHA
38	5914418184422	5580234	F	83	JARDIM DA PENHA

How?

Select Age column > Click Home tab > goto Editing tab > Click Find & Select > Go To Special > Constants > Uncheck Numbers > OK

Remove Identified ?

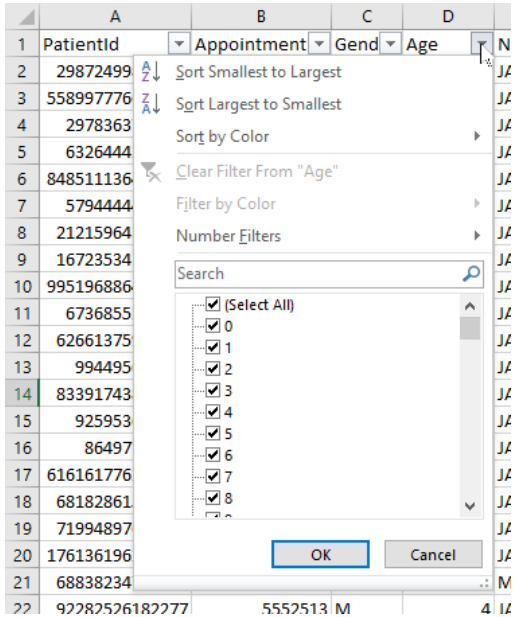


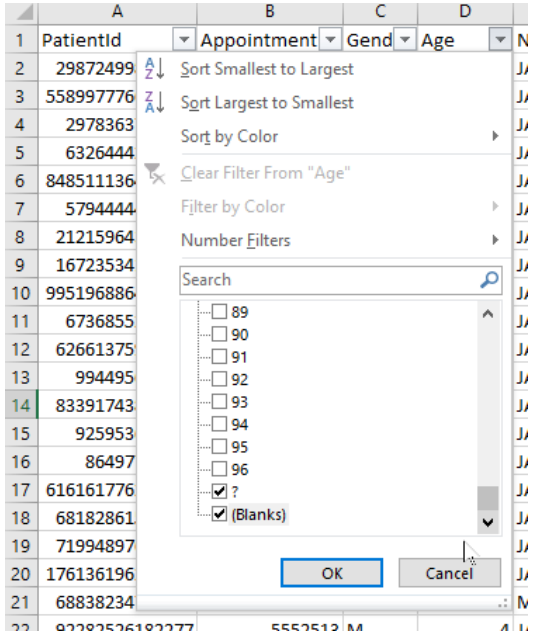


How?

Select row with the identified ? or Blank > Right click Delete

Can Also Use Filter to Identify Missing Data



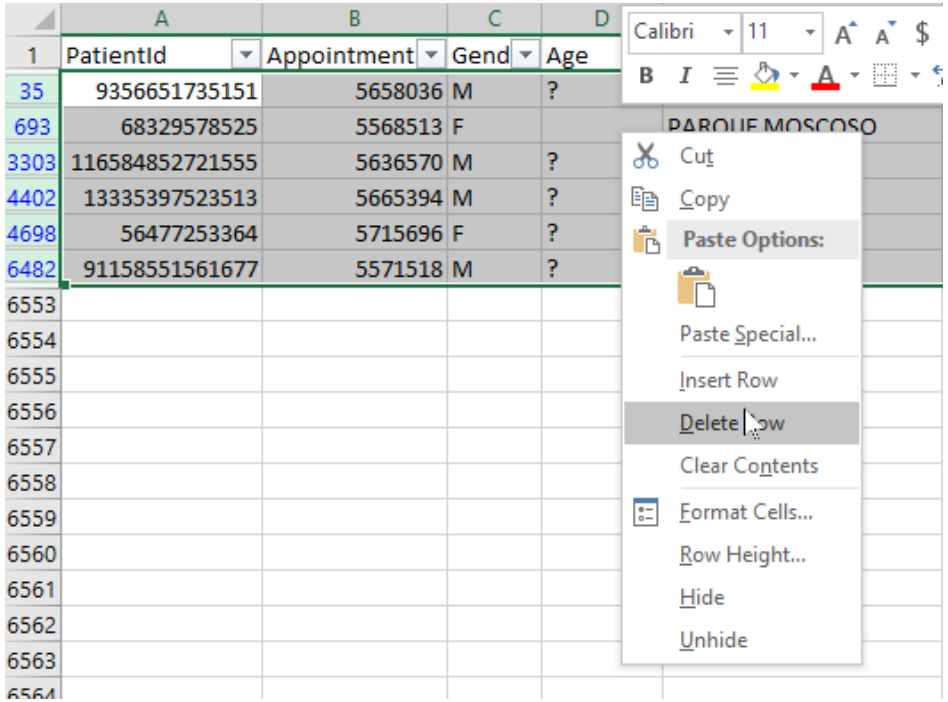


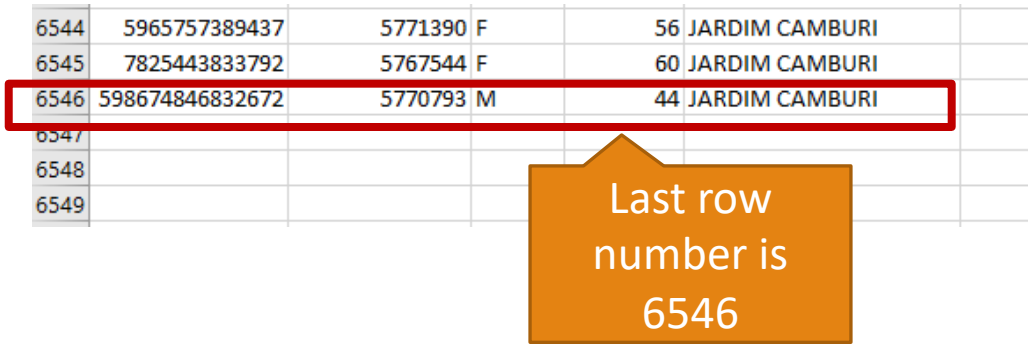
	PatientId	Appointment	Gend	Age	Neighbourhood
35	9356651735151	5658036	M	?	JARDIM DA PENHA
693	68329578525	5568513	F		PARQUE MOSCOSO
3303	116584852721555	5636570	M	?	MARIA ORTIZ
4402	13335397523513	5665394	M	?	CONQUISTA
4698	56477253364	5715696	F	?	DA PENHA
6482	91158551561677	5571518	M	?	JARDIM CAMBURI

How?

Click any populated area > Click Data tab > goto Sort & Filter > Click Filter > Click Age column down arrow > uncheck Select All > scroll down > check ? and Blanks

Remove Filtered Rows





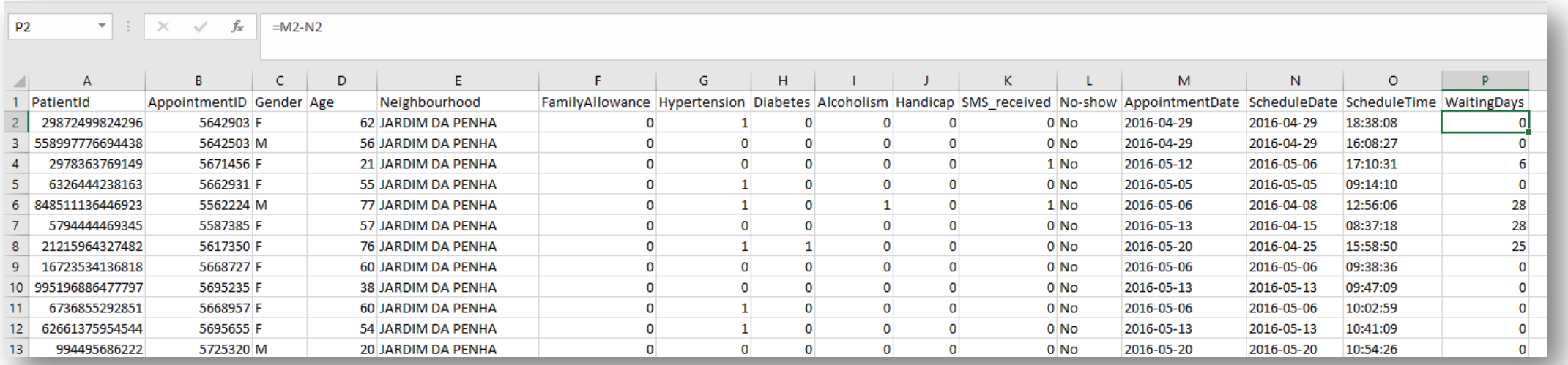
How?

Select rows > Right click > Delete Row

REMOVE CONFLICTING DATA

WaitingDays Cannot be Less Than 0

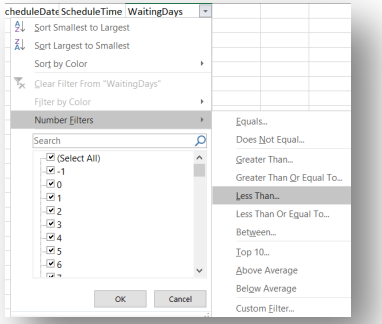
Add a new column WaitingDays.
WaitingDays = AppointmentDate - ScheduleDate
The number of days patient waited:
0 day – ScheduleDate and AppointmentDate are on the same day
1 day – AppointmentDate happens one day after ScheduleDate

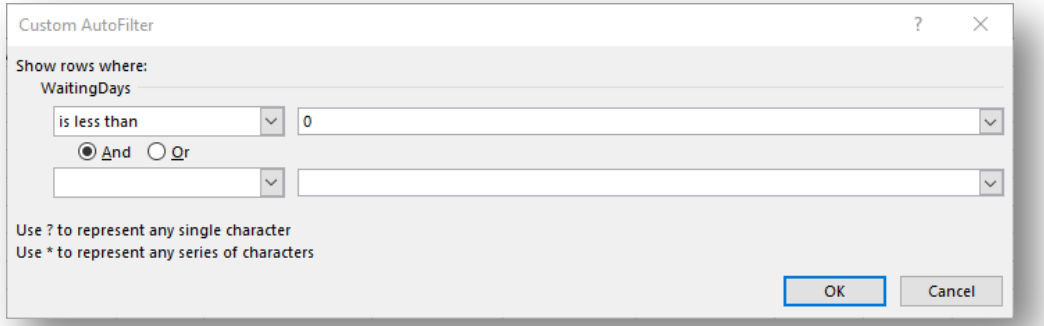


P2 | =M2-N2

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	PatientId	AppointmentID	Gender	Age	Neighbourhood	FamilyAllowance	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	No-show	AppointmentDate	ScheduleDate	ScheduleTime	WaitingDays
2	29872499824296	5642903	F	62	JARDIM DA PENHA	0	1	0	0	0	0	No	2016-04-29	2016-04-29	18:38:08	0
3	558997776694438	5642503	M	56	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-04-29	2016-04-29	16:08:27	0
4	2978363769149	5671456	F	21	JARDIM DA PENHA	0	0	0	0	0	1	No	2016-05-12	2016-05-06	17:10:31	6
5	6326444238163	5662931	F	55	JARDIM DA PENHA	0	1	0	0	0	0	No	2016-05-05	2016-05-05	09:14:10	0
6	848511136446923	5562224	M	77	JARDIM DA PENHA	0	1	0	1	0	1	No	2016-05-06	2016-04-08	12:56:06	28
7	5794444469345	5587385	F	57	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-13	2016-04-15	08:37:18	28
8	21215964327482	5617350	F	76	JARDIM DA PENHA	0	1	1	0	0	0	No	2016-05-20	2016-04-25	15:58:50	25
9	16723534136818	5668727	F	60	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-06	2016-05-06	09:38:36	0
10	995196886477797	5695235	F	38	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-13	2016-05-13	09:47:09	0
11	6736855292851	5668957	F	60	JARDIM DA PENHA	0	1	0	0	0	0	No	2016-05-06	2016-05-06	10:02:59	0
12	62661375954544	5695655	F	54	JARDIM DA PENHA	0	1	0	0	0	0	No	2016-05-13	2016-05-13	10:41:09	0
13	994495686222	5725320	M	20	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-20	2016-05-20	10:54:26	0

Filter WaitingDays >= 0





	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	PatientId	Appointment	Gend	Age	Neighbourhood	FamilyAllowan	Hypertensio	Diabet	Alcoholis	Handic	SMS_receive	No-sho	AppointmentDa	ScheduleDa	ScheduleTin	WaitingDays
113	211759384191	5650729	M	18	JARDIM DA PENHA	0	0	0	0	0	0	No	2016-05-03	2016-05-04	08:03:29	-1
3435	5587789811398	5726267	M	14	SANTA TEREZA	0	0	0	0	0	0	Yes	2016-05-19	2016-05-20	13:24:18	-1
6547																
6542	5965757389437	5771390	F	56	JARDIM CAMBURI	0	0	0	0	0	1	No	2016-06-07	2016-06-03	12:45:35	4
6544	598674846832672	5770793	M	44	JARDIM CAMBURI	0	0	0	0	0	1	No	2016-06-07	2016-06-03	10:57:50	4
6545																
6546																

Last row number is 6544

How?

Click any populated area > Click Data tab > goto Sort & Filter > Click Filter > Click WaitingDays column down arrow > Select Number Filters > Less Than > Enter 0 > OK

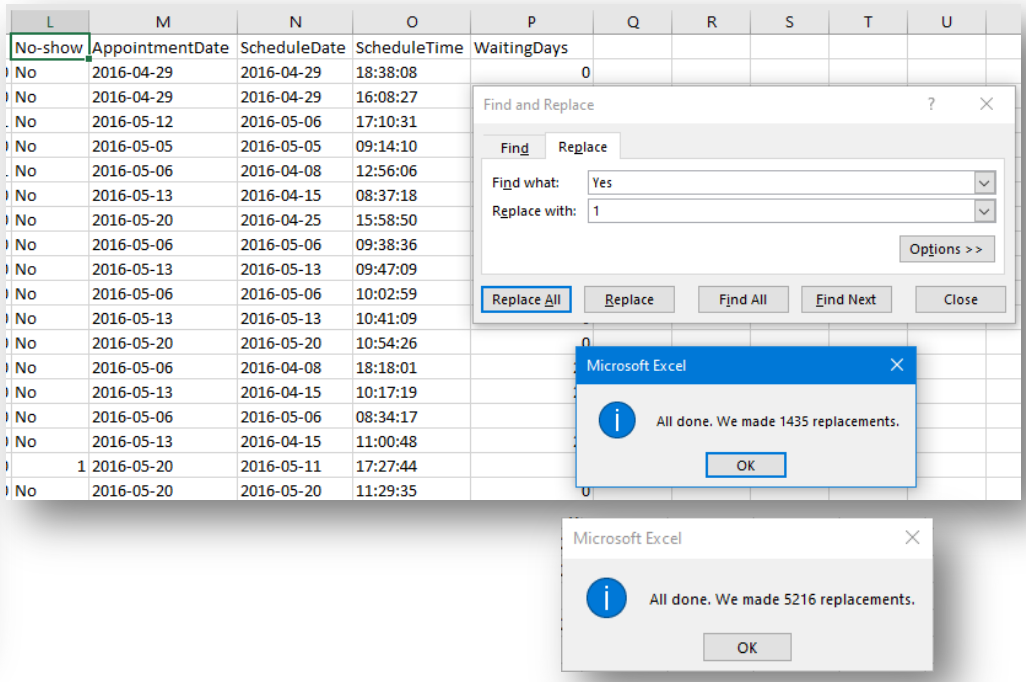
Select both rows > Right click Delete Row

FIND AND REPLACE TEXTS

Replace Text with Binary Data

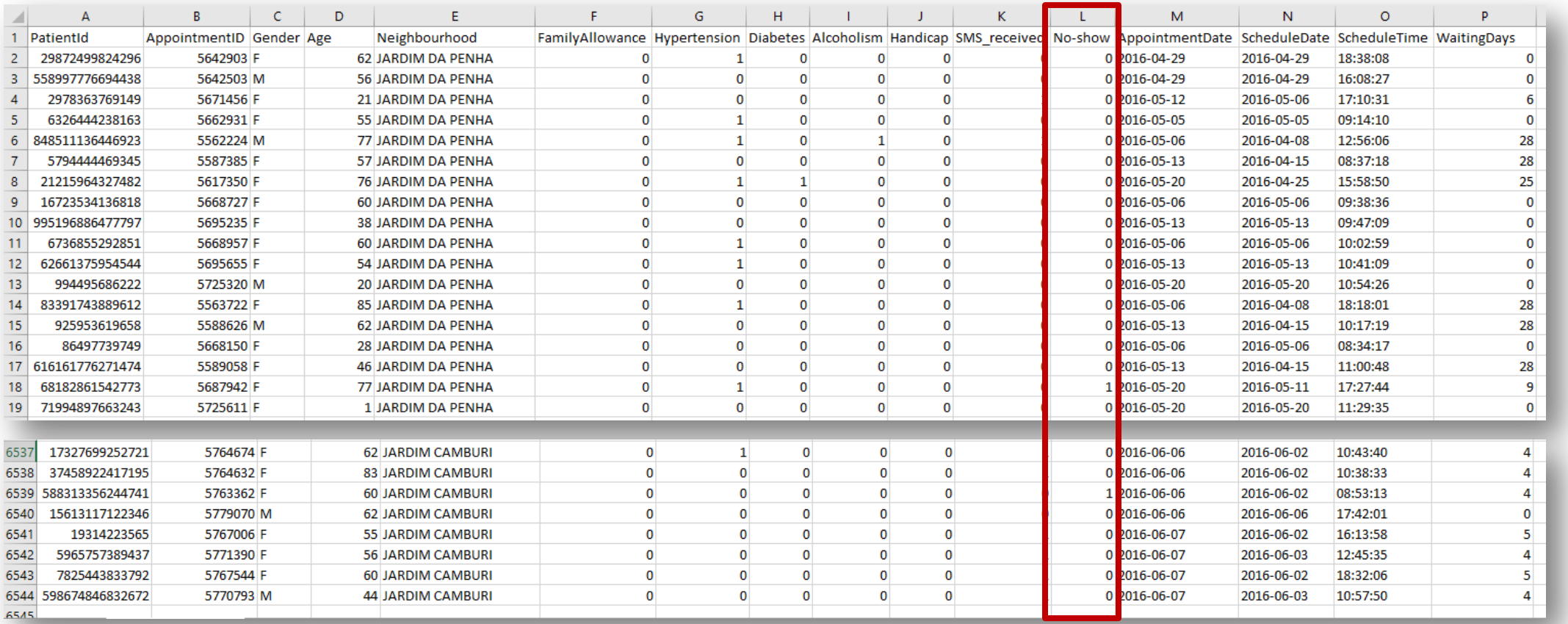
For convenience of analysis:
Yes – 1
No – 0

J	K	L	M
Handicap	SMS_received	No-show	AppointmentDate
0	0	No	2016-04-29
0	0	No	2016-04-29
0	1	No	2016-05-12
0	0	No	2016-05-05
0	1	No	2016-05-06
0	0	No	2016-05-13
0	0	No	2016-05-20
0	0	No	2016-05-06
0	0	No	2016-05-13
0	0	No	2016-05-06
0	0	No	2016-05-13
0	0	No	2016-05-20
0	0	No	2016-05-06
0	0	No	2016-05-13
0	0	No	2016-05-06
0	0	No	2016-05-13
0	0	Yes	2016-05-20
0	0	No	2016-05-20
0	0	No	2016-05-05



How?
Click any populated area > Click Home tab > goto Editing tab > Click Find & Select > Replace > Enter “Yes” > “1” > Click Replace All > OK
Enter “No” > “0” > Click Replace All > OK

So Far,



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	PatientId	AppointmentID	Gender	Age	Neighbourhood	FamilyAllowance	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	No-show	AppointmentDate	ScheduleDate	ScheduleTime	WaitingDays
2	29872499824296	5642903	F	62	JARDIM DA PENHA	0	1	0	0	0		0	2016-04-29	2016-04-29	18:38:08	0
3	558997776694438	5642503	M	56	JARDIM DA PENHA	0	0	0	0	0		0	2016-04-29	2016-04-29	16:08:27	0
4	2978363769149	5671456	F	21	JARDIM DA PENHA	0	0	0	0	0		0	2016-05-12	2016-05-06	17:10:31	6
5	6326444238163	5662931	F	55	JARDIM DA PENHA	0	1	0	0	0		0	2016-05-05	2016-05-05	09:14:10	0
6	848511136446923	5562224	M	77	JARDIM DA PENHA	0	1	0	1	0		0	2016-05-06	2016-04-08	12:56:06	28
7	5794444469345	5587385	F	57	JARDIM DA PENHA	0	0	0	0	0		0	2016-05-13	2016-04-15	08:37:18	28
8	21215964327482	5617350	F	76	JARDIM DA PENHA	0	1	1	0	0		0	2016-05-20	2016-04-25	15:58:50	25
9	16723534136818	5668727	F	60	JARDIM DA PENHA	0	0	0	0	0		0	2016-05-06	2016-05-06	09:38:36	0
10	995196886477797	5695235	F	38	JARDIM DA PENHA	0	0	0	0	0		0	2016-05-13	2016-05-13	09:47:09	0
11	6736855292851	5668957	F	60	JARDIM DA PENHA	0	1	0	0	0		0	2016-05-06	2016-05-06	10:02:59	0
12	62661375954544	5695655	F	54	JARDIM DA PENHA	0	1	0	0	0		0	2016-05-13	2016-05-13	10:41:09	0
13	994495686222	5725320	M	20	JARDIM DA PENHA	0	0	0	0	0		0	2016-05-20	2016-05-20	10:54:26	0
14	83391743889612	5563722	F	85	JARDIM DA PENHA	0	1	0	0	0		0	2016-05-06	2016-04-08	18:18:01	28
15	925953619658	5588626	M	62	JARDIM DA PENHA	0	0	0	0	0		0	2016-05-13	2016-04-15	10:17:19	28
16	86497739749	5668150	F	28	JARDIM DA PENHA	0	0	0	0	0		0	2016-05-06	2016-05-06	08:34:17	0
17	616161776271474	5589058	F	46	JARDIM DA PENHA	0	0	0	0	0		0	2016-05-13	2016-04-15	11:00:48	28
18	68182861542773	5687942	F	77	JARDIM DA PENHA	0	1	0	0	0		1	2016-05-20	2016-05-11	17:27:44	9
19	71994897663243	5725611	F	1	JARDIM DA PENHA	0	0	0	0	0		0	2016-05-20	2016-05-20	11:29:35	0
6537	17327699252721	5764674	F	62	JARDIM CAMBURI	0	1	0	0	0		0	2016-06-06	2016-06-02	10:43:40	4
6538	37458922417195	5764632	F	83	JARDIM CAMBURI	0	0	0	0	0		0	2016-06-06	2016-06-02	10:38:33	4
6539	588313356244741	5763362	F	60	JARDIM CAMBURI	0	0	0	0	0		1	2016-06-06	2016-06-02	08:53:13	4
6540	15613117122346	5779070	M	62	JARDIM CAMBURI	0	0	0	0	0		0	2016-06-06	2016-06-06	17:42:01	0
6541	19314223565	5767006	F	55	JARDIM CAMBURI	0	0	0	0	0		0	2016-06-07	2016-06-02	16:13:58	5
6542	5965757389437	5771390	F	56	JARDIM CAMBURI	0	0	0	0	0		0	2016-06-07	2016-06-03	12:45:35	4
6543	7825443833792	5767544	F	60	JARDIM CAMBURI	0	0	0	0	0		0	2016-06-07	2016-06-02	18:32:06	5
6544	598674846832672	5770793	M	44	JARDIM CAMBURI	0	0	0	0	0		0	2016-06-07	2016-06-03	10:57:50	4

BIN DATA INTO BUCKETS

Bin Age into Buckets

Add a new column AgeGroup.
AgeGroup =VLOOKUP(Age, $R$7:$S$11, 2, 1)
Child: 0 – 12
Adolescence: 13 – 18
YoungAdult: 19 – 45
MiddleAge: 46 – 60
Senior: >60

True = Approximate match

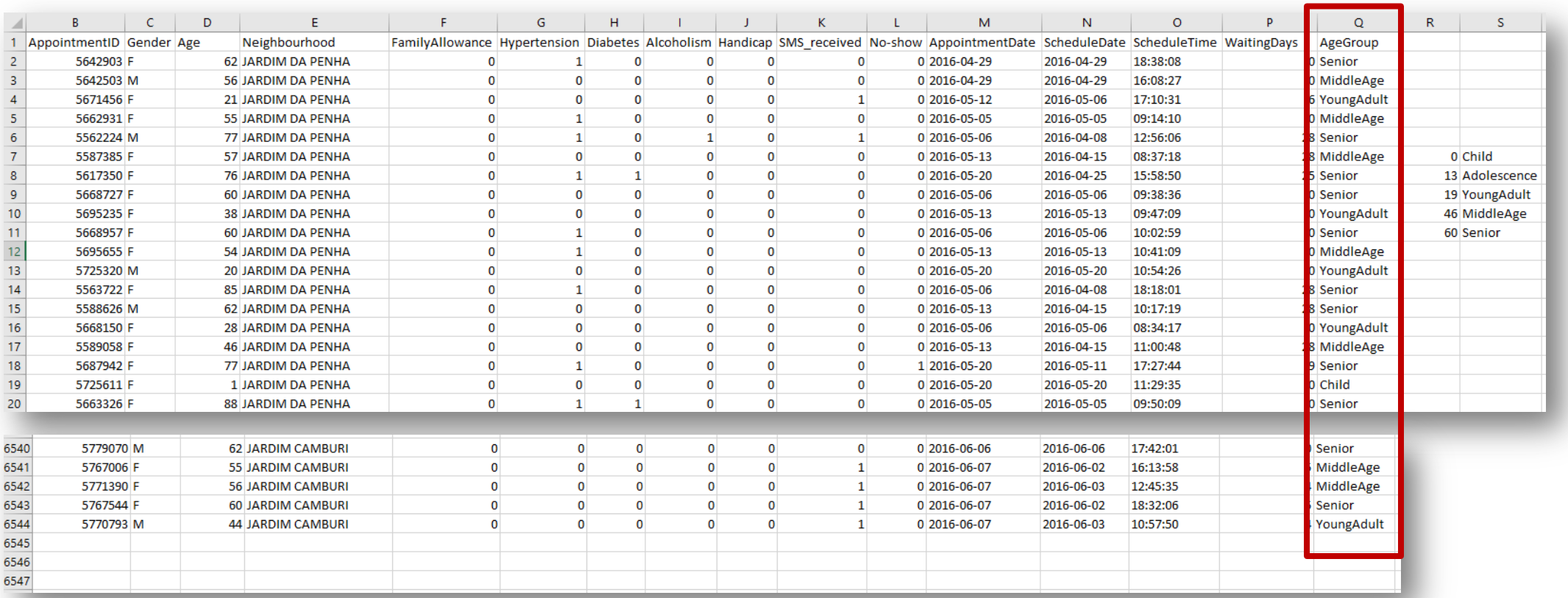
Q	R	S
AgeGroup		
Senior		
MiddleAge		
YoungAdult		
MiddleAge		
Senior		
MiddleAge	0	Child
Senior	13	Adolescence
Senior	19	YoungAdult
YoungAdult	46	MiddleAge
Senior	60	Senior

=VLOOKUP(D2, $R$7:$S$11, 2, 1)

D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Age	Neighbourhood	FamilyAllowance	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	No-show	AppointmentDate	ScheduleDate	ScheduleTime	WaitingDays	AgeGroup		
62	JARDIM DA PENHA	0	1	0	0	0	0	0	2016-04-29	2016-04-29	18:38:08	0	Senior		
56	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-04-29	2016-04-29	16:08:27	0	MiddleAge		
21	JARDIM DA PENHA	0	0	0	0	0	1	0	2016-05-12	2016-05-06	17:10:31	6	YoungAdult		
55	JARDIM DA PENHA	0	1	0	0	0	0	0	2016-05-05	2016-05-05	09:14:10	0	MiddleAge		
77	JARDIM DA PENHA	0	1	0	1	0	1	0	2016-05-06	2016-04-08	12:56:06	28	Senior		
57	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-13	2016-04-15	08:37:18	28	MiddleAge	0	Child
76	JARDIM DA PENHA	0	1	1	0	0	0	0	2016-05-20	2016-04-25	15:58:50	25	Senior	13	Adolescence
60	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-06	2016-05-06	09:38:36	0	Senior	19	YoungAdult
38	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-13	2016-05-13	09:47:09	0	YoungAdult	46	MiddleAge
60	JARDIM DA PENHA	0	1	0	0	0	0	0	2016-05-06	2016-05-06	10:02:59	0	Senior	60	Senior
54	JARDIM DA PENHA	0	1	0	0	0	0	0	2016-05-13	2016-05-13	10:41:09	0	MiddleAge		

2nd column

So Far,



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	AppointmentID	Gender	Age	Neighbourhood	FamilyAllowance	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	No-show	AppointmentDate	ScheduleDate	ScheduleTime	WaitingDays	AgeGroup		
2	5642903	F	62	JARDIM DA PENHA	0	1	0	0	0	0	0	2016-04-29	2016-04-29	18:38:08	0	Senior		
3	5642503	M	56	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-04-29	2016-04-29	16:08:27	0	MiddleAge		
4	5671456	F	21	JARDIM DA PENHA	0	0	0	0	0	1	0	2016-05-12	2016-05-06	17:10:31	6	YoungAdult		
5	5662931	F	55	JARDIM DA PENHA	0	1	0	0	0	0	0	2016-05-05	2016-05-05	09:14:10	0	MiddleAge		
6	5562224	M	77	JARDIM DA PENHA	0	1	0	1	0	1	0	2016-05-06	2016-04-08	12:56:06	28	Senior		
7	5587385	F	57	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-13	2016-04-15	08:37:18	28	MiddleAge	0	Child
8	5617350	F	76	JARDIM DA PENHA	0	1	1	0	0	0	0	2016-05-20	2016-04-25	15:58:50	25	Senior	13	Adolescence
9	5668727	F	60	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-06	2016-05-06	09:38:36	0	Senior	19	YoungAdult
10	5695235	F	38	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-13	2016-05-13	09:47:09	0	YoungAdult	46	MiddleAge
11	5668957	F	60	JARDIM DA PENHA	0	1	0	0	0	0	0	2016-05-06	2016-05-06	10:02:59	0	Senior	60	Senior
12	5695655	F	54	JARDIM DA PENHA	0	1	0	0	0	0	0	2016-05-13	2016-05-13	10:41:09	0	MiddleAge		
13	5725320	M	20	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-20	2016-05-20	10:54:26	0	YoungAdult		
14	5563722	F	85	JARDIM DA PENHA	0	1	0	0	0	0	0	2016-05-06	2016-04-08	18:18:01	28	Senior		
15	5588626	M	62	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-13	2016-04-15	10:17:19	28	Senior		
16	5668150	F	28	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-06	2016-05-06	08:34:17	0	YoungAdult		
17	5589058	F	46	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-13	2016-04-15	11:00:48	28	MiddleAge		
18	5687942	F	77	JARDIM DA PENHA	0	1	0	0	0	0	1	2016-05-20	2016-05-11	17:27:44	9	Senior		
19	5725611	F	1	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-20	2016-05-20	11:29:35	0	Child		
20	5663326	F	88	JARDIM DA PENHA	0	1	1	0	0	0	0	2016-05-05	2016-05-05	09:50:09	0	Senior		
6540	5779070	M	62	JARDIM CAMBURI	0	0	0	0	0	0	0	2016-06-06	2016-06-06	17:42:01	0	Senior		
6541	5767006	F	55	JARDIM CAMBURI	0	0	0	0	0	1	0	2016-06-07	2016-06-02	16:13:58	5	MiddleAge		
6542	5771390	F	56	JARDIM CAMBURI	0	0	0	0	0	1	0	2016-06-07	2016-06-03	12:45:35	4	MiddleAge		
6543	5767544	F	60	JARDIM CAMBURI	0	0	0	0	0	1	0	2016-06-07	2016-06-02	18:32:06	5	Senior		
6544	5770793	M	44	JARDIM CAMBURI	0	0	0	0	0	1	0	2016-06-07	2016-06-03	10:57:50	4	YoungAdult		
6545																		
6546																		
6547																		

FIND OUTLIERS

Outliers in WaitingDays

Calculate Quartiles
- Q1 =QUARTILE(P1:P6544, 1)
- Q3 =QUARTILE(P1:P6544, 3)

Calculate Upper and Lower Boundaries
- IQR = Q3 – Q1
- L BOUND = Q1 – IQR * 1.5
- U BOUND = Q3 + IQR * 1.5

Outlier = WaitingDays < LB or WaitingDays > UB

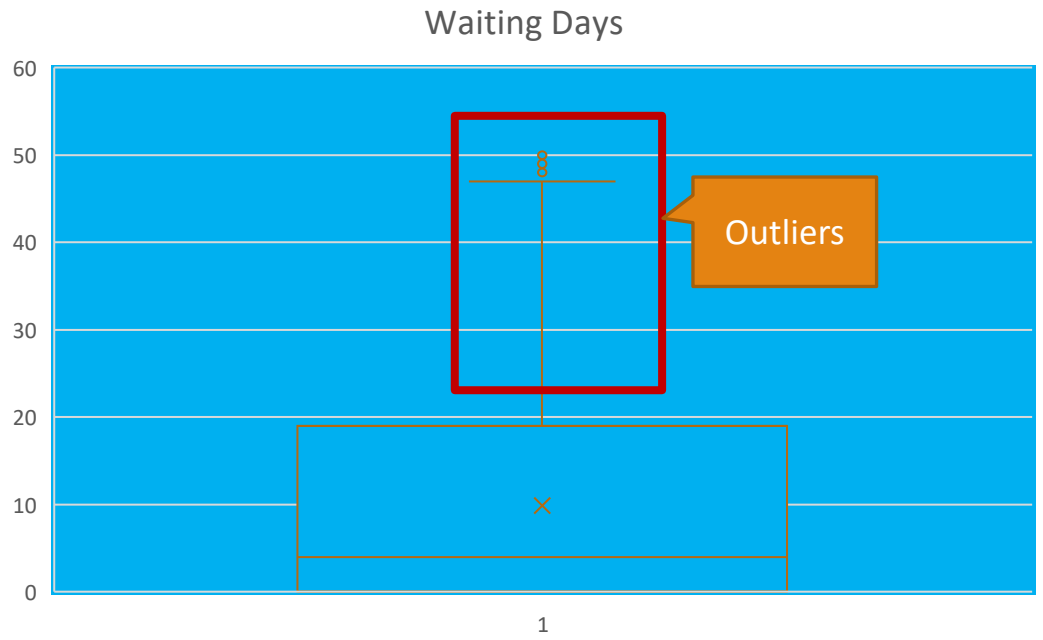
There are 96 outliers (revealed with filtering TRUE)

P	Q	R	S	T
WaitingDays	Outlier	AgeGroup		
0		Senior		
0		MiddleAge		
6		YoungAdult		
0		MiddleAge		
28		Senior		
28		MiddleAge	0	Child
25		Senior	13	Adolescence
0		Senior	19	YoungAdult
0		YoungAdult	46	MiddleAge
0		Senior	60	Senior
0		MiddleAge		
0		YoungAdult		
28		Senior	Q1	0
28		Senior	Q3	20
0		YoungAdult	IQR	20
28		MiddleAge	L BOUND	-30
9		Senior	U BOUND	50
0		Child		

=OR(P2<$T$17, P2>$T$18)

E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Neighbourhood	FamilyAllowance	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	No-show	AppointmentDate	ScheduleDate	ScheduleTime	WaitingDays	Outlier	AgeGroup
JARDIM DA PENHA	0	1	0	0	0	0	0	2016-04-29	2016-04-29	18:38:08	0	FALSE	Senior
JARDIM DA PENHA	0	0	0	0	0	0	0	2016-04-29	2016-04-29	16:08:27	0		MiddleAge
JARDIM DA PENHA	0	0	0	0	0	1	0	2016-05-12	2016-05-06	17:10:31	6		YoungAdult
JARDIM DA PENHA	0	1	0	0	0	0	0	2016-05-05	2016-05-05	09:14:10	0		MiddleAge

Produce Box and Whisker Plot

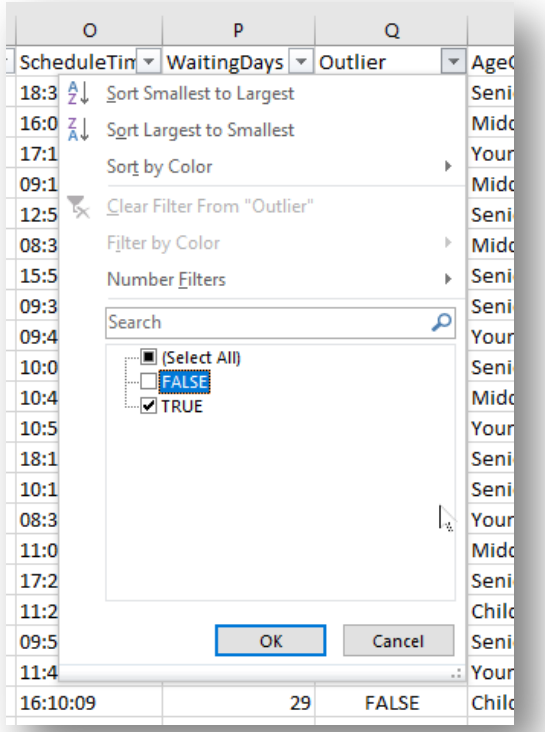


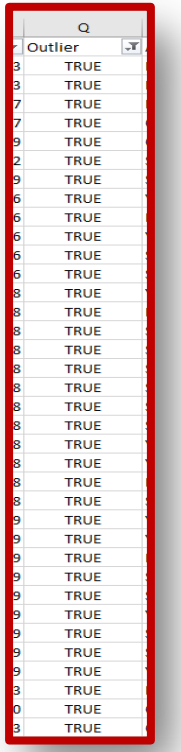
How?

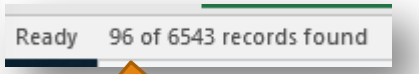
Select WaitingDays column > Click Insert tab > goto Charts tab > Click Recommended Charts > Select All Charts > Box and Whisker > OK > Change Chart Title to “Waiting Days”

Filter Outliers in WaitingDays

There are 96 outliers
(revealed with filtering TRUE)





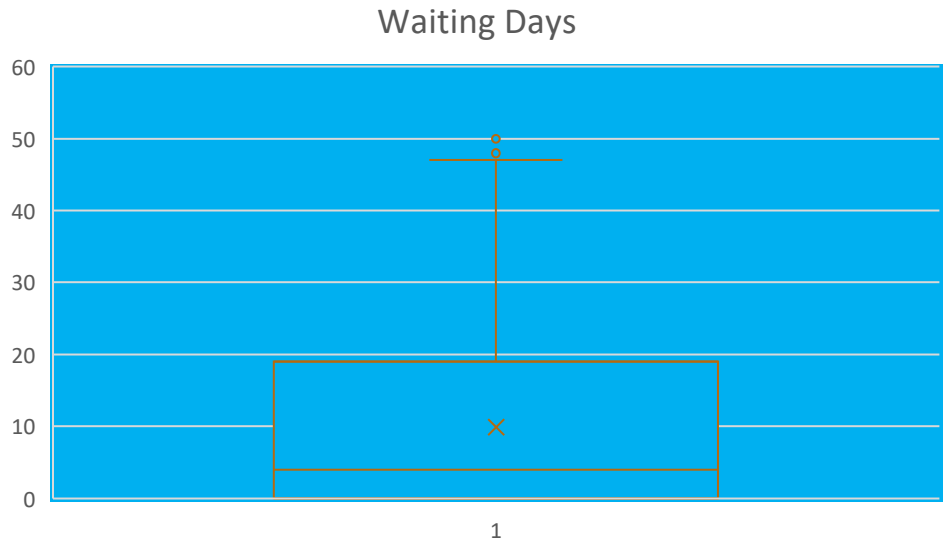


Delete all 96 rows.

6445	19314223565	5767006	F	55	JARDIM CAMBURI	0	0	0	0	0	1	0	2016-06-07	2016-06-02	16:13:58	5	FALSE	MiddleAge
6446	5965757389437	5771390	F	56	JARDIM CAMBURI	0	0	0	0	0	1	0	2016-06-07	2016-06-03	12:45:35	4	FALSE	MiddleAge
6447	7825443833792	5767544	F	60	JARDIM CAMBURI	0	0	0	0	0	1	0	2016-06-07	2016-06-02	18:32:06	5	FALSE	Senior
6448	598674846832672	5770793	M	44	JARDIM CAMBURI	0	0	0	0	0	1	0	2016-06-07	2016-06-03	10:57:50	4	FALSE	YoungAdult
6449																		
6450																		

Last row number is 6448

Produce Box and Whisker Plot After Deleting Outliers



How?

Select WaitingDays column > Click Insert tab > goto Charts tab > Click Recommended Charts > Select All Charts > Box and Whisker > OK > Change Chart Title to “Waiting Days”

DATA CLEANING

DR DANNY POO

BIG DATA ANALYTICS AND VISUALISATION