

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	0	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	0	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	0	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	0	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	0	No

Initial insights:
6573 rows, 14 columns

CHANGE DATA FORMAT

Scientific Data Format for PatientId

Unsuitable: Change to Text Format ..1/3

The image shows two side-by-side tables in Microsoft Excel. The left table, titled 'PatientId', contains 20 rows of scientific notation values. The right table, titled 'AppointmentID', contains 22 rows of integer values. Between them is a red vertical bar. To the right of the tables is a screenshot of the 'Convert Text to Columns Wizard - Step 1 of 3' dialog box. The dialog box has several sections: 'Original data type' (set to 'Delimited'), 'Choose the file type that best describes your data' (radio button selected for 'Delimited'), 'Preview of selected data:' (showing the first six rows of the 'PatientId' column), and buttons for 'Cancel', '< Back', 'Next >', and 'Finish'. A large orange callout box at the bottom right contains the text 'How?' and 'Click Data tab > Text to Columns > Delimited > Next'.

	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	C	D	E	F	G	H
1	PatientId	AppointmentID		Convert Text to Columns Wizard - Step 1 of 3				
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

The screenshot shows a Microsoft Excel spreadsheet with a column of Patient IDs in scientific notation (e.g., 2.98725E+13) in column A. To its right is the 'Convert Text to Columns Wizard - Step 2 of 3' dialog box. The dialog box allows setting delimiters for the text. Under 'Delimiters', none are selected. Under 'Text qualifier', a single quote character is chosen. The 'Data preview' section shows the original data, which appears as plain text due to the lack of a delimiter. At the bottom, the 'Next >' button is highlighted in blue.

How?
Click Next

Scientific Data Format for PatientId

Unsuitable: Change to Text Format ..3/3

The image shows three panels illustrating the conversion process:

- Panel 1 (Left):** A screenshot of an Excel spreadsheet with a single column labeled "PatientId". The first few rows contain values like 2.98725E+13, 5.58998E+14, etc., in scientific notation.
- Panel 2 (Middle):** A screenshot of the "Convert Text to Columns Wizard - Step 3 of 3" dialog box. The "Column data format" section has the "Text" radio button selected. The "Destination" field is set to "\$A\$1". The "Data preview" section shows the PatientId values converted to plain text. The "Finish" button is highlighted.
- Panel 3 (Right):** A screenshot of the Excel spreadsheet after the conversion. The "PatientId" column now contains the values as plain text (e.g., 29872499824296, 558997776694438, etc.) instead of scientific notation.

How?
Select Text > Finish

Scientific Data Format for PatientId

Unsuitable: Change to Number Format

The screenshot shows a Microsoft Excel spreadsheet with a single column labeled "PatientId". The first few rows contain large numbers in scientific notation, such as 2.98725E+13, 5.58998E+14, and 2.97836E+12. A context menu is open over the second row, with "Format Cells..." selected. The "Format Cells" dialog box is displayed, showing the "Number" tab selected. In the "Category" list, "Number" is highlighted. The "Sample" box shows "PatientId" with a value of 29872499824296. The "Decimal places:" dropdown is set to 0. Below it, the "Negative numbers:" section shows a list of four formats: -1234, 1234, -1234, and -1234. To the right of the dialog box, the full column of PatientId values is shown in standard numerical 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 29872499824296
3 558997776694438
4 2978363769149
5 6326444238163
6 55583448227198
7 848511136446923
8 579444469345
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
2016-04-23T15:58:50Z	2016-05-20T00:00:00Z
2016-05-06T09:38:36Z	2016-05-06T00:00:00Z
2016-05-07T10:02:59Z	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
2016-04-13T11:00:48Z	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
2016-04-19T15:38:24Z	2016-05-18T00:00:00Z
2016-05-02T09:52:17Z	2016-05-04T00:00:00Z
2016-05-13T15:49:02Z	2016-05-12T00:00:00Z
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	J
1	PatientId	AppointmentID	Gender	ScheduleDay	ScheduleDate	ScheduleTime	AppointmentDay	AppointmentDate	Age	DOB
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

	A	B	C	D	E	F
1	PatientId	AppointmentId	Gender	ScheduleDay	ScheduleDate	ScheduleTime
2	29872499824296	5642903	F	2016-04-29T18:38:08Z	2016-04-29	
3	558997776694438	5642503	M	2016-04-29T16:08:27Z	2016-04-29	

Use Split functions: LEFT, MID

	A	B	C	D	E	F
1	PatientId	AppointmentId	Gender	ScheduleDay	ScheduleDate	ScheduleTime
2	29872499824296	5642903	F	2016-04-29T18:38:08Z	2016-04-29	18:38:08
3	558997776694438	5642503	M	2016-04-29T16:08:27Z	2016-04-29	16:08:27

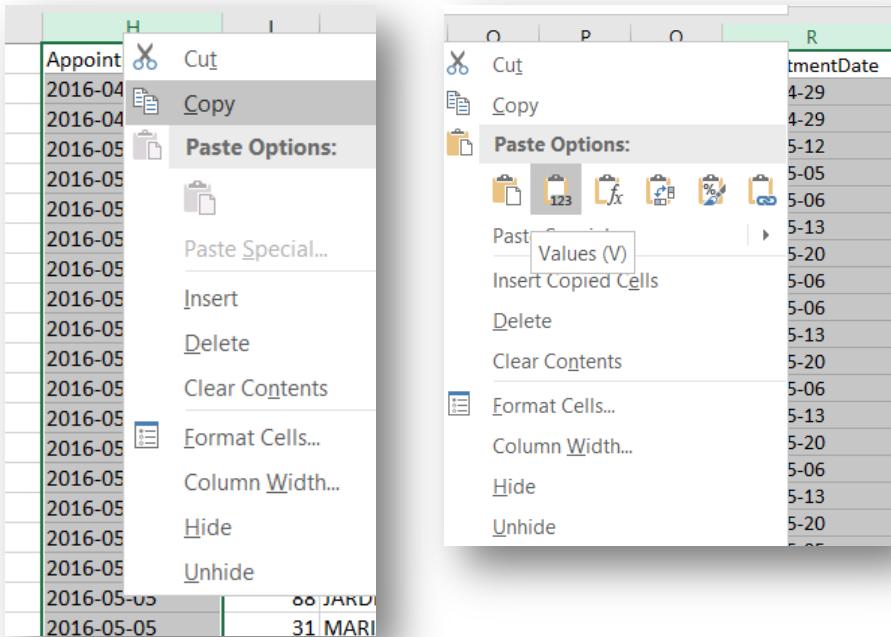
ScheduleDate
 $=LEFT(ScheduleDay, 10)$

	E	F	G	H	I
	ScheduleDate	ScheduleTime	AppointmentDay	AppointmentDate	Age
38:08Z	2016-04-29	18:38:08	2016-04-29T00:00:00Z	2016-04-29	
08:27Z	2016-04-29	16:08:27	2016-04-29T00:00:00Z	2016-04-29	
10:31Z	2016-05-06	17:10:31	2016-05-12T00:00:00Z	2016-05-12	

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

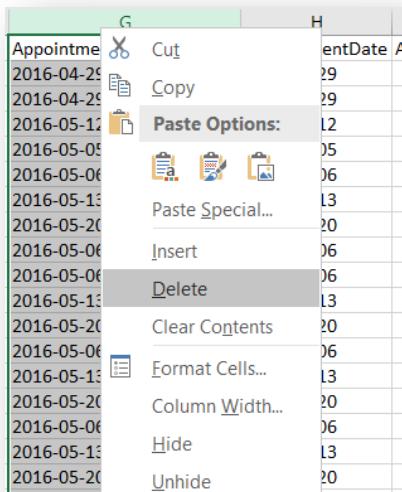
How?

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

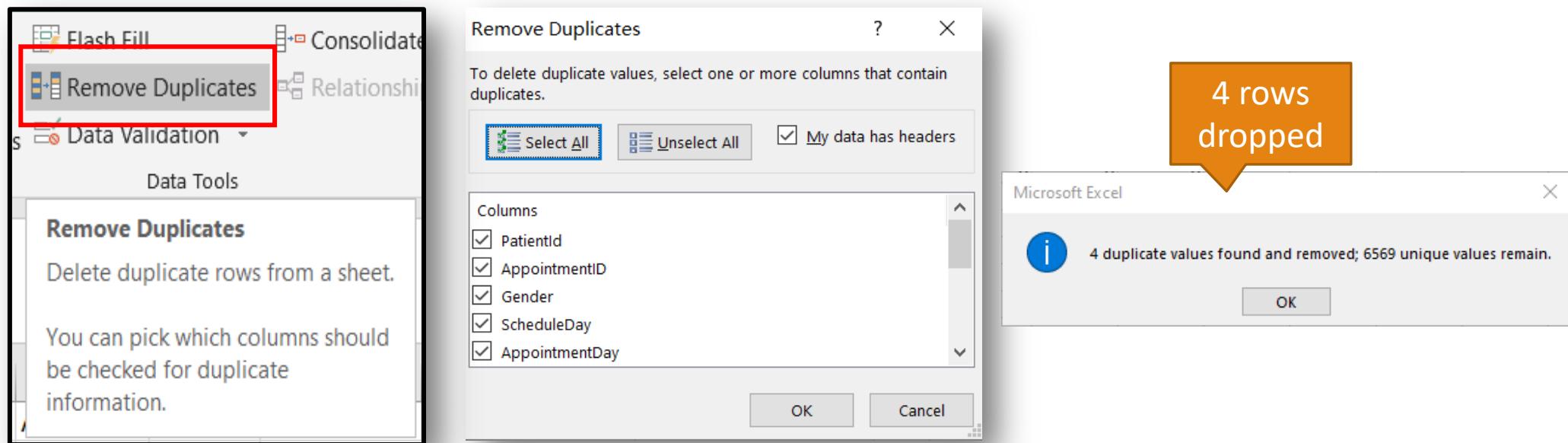
M	N	O
AppointmentDate	ScheduleDate	ScheduleTime
2016-04-29	2016-04-29	18:38:08
2016-04-29	2016-04-29	16:08:27
2016-05-12	2016-05-06	17:10:31
2016-05-05	2016-05-05	09:14:10
2016-05-12	2016-04-20	13:12:17
2016-05-06	2016-04-08	12:56:06
2016-05-13	2016-04-15	08:37:18
2016-05-20	2016-04-25	15:58:50
2016-05-06	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

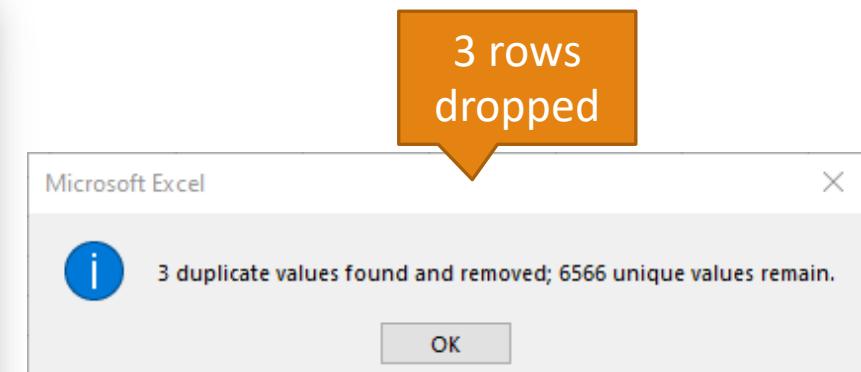


How?

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

AppointmentID Cannot be Repeated

	A	B	C	D	E
1	PatientId	AppointmentID	Gender	ScheduleDay	AppointmentDay
2	29872499824296	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z
3	558997776694438	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z
4	2978363769149	5671456	F	2016-05-06T17:10:31Z	2016-05-12T00:00:00Z
5	6326444238163	56629			
6	55583448227198	56080			
7	848511136446923	55622			
8	5794444469345	55873			
9	21215964327482	56173			
10	16723534136818	56687			
11	995196886477797	56952			
12	6736855292851	56689			
13	62661375954544	56956			
14	994495686222	57253			
15	83391743889612	55637			
16	925953619658	55886			
17	52983244522738	57112			
18	86497739749	56681			
19	616161776271171	55800			

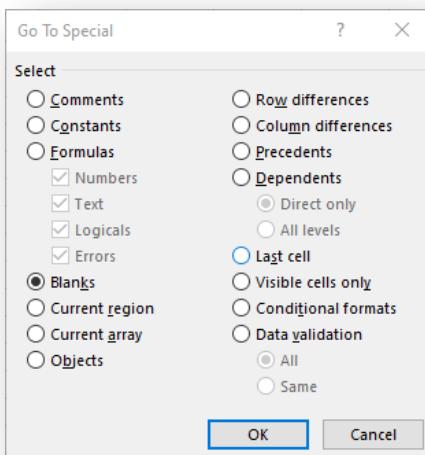


How?

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	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

The screenshot shows two parts of the Microsoft Excel interface. On the left, a context menu is open over a blank cell in row 17, column A. The menu includes options like Cut, Copy, Paste Options, Paste Special, Smart Lookup, Insert, Delete (which is highlighted), Clear Contents, Quick Analysis, Filter, Sort, Insert Comment, Delete Comment, Format Cells, Pick From Drop-down List, Define Name, and Hyperlink. On the right, a 'Delete' dialog box is displayed over a table. It contains four radio button options: Shift cells left, Shift cells up, Entire row (which is selected), and Entire column. Below the radio buttons are OK and Cancel buttons.

A	B	C	D	E
55583448227198	5608066 M	JARDIM DA PENHA		
848511136446923	5562224 M			
5794444469345	5587385 F			
21215964327482	5617350 F			
16723534136818	5668727 F			
995196886477797	5695235 F			
6736855292851	5668957 F			
62661375954544	5695655 F			
994495686222	5725320 M			
83391743889612	5563722 F			
925953619658	5588626 M			
52983244522738	5711227 F			
86497739749	5668150 F			
616161776271474	5589058 F			
68182861542773	5687942 F			
7199487663243	5725611 F			
176136196269528	5663326 F			
68838234725772	5664257 M			
92282526182277	5552513 M			
525313675286266	5608074 F			
49143376656174	5550941 F			
13726233339	5558708 M			
623118677783	5579128 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

The screenshot shows a Microsoft Excel spreadsheet with columns labeled D, E, F, and G. The first row contains headers: 'Age', 'Neighbourhood', 'FamilyAllowance', and 'Hypertension'. Below the headers, there are several rows of data. A blue rectangular selection highlights the first few rows of data. A 'Go To Special' dialog box is overlaid on the spreadsheet. The 'Select' tab is active, showing various options like 'Comments', 'Constants', 'Formulas', etc. The 'Numbers' checkbox is checked, while 'Text' and 'Logicals' are also checked. The 'Errors' checkbox is checked. The 'Blanks' checkbox is checked. The 'Current region' checkbox is checked. The 'Current array' checkbox is checked. The 'Objects' checkbox is checked. The 'OK' button is highlighted with a blue border.

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 ?

35	9356651735151	5658036	M	?
36	21833136218142	5653210	F	
37	31277536312512	5561250	F	
38	5914418184422	5580234	F	
39	297793143163734	5602465	M	
40	743733582334942	5602969	F	
41	17385944962	5624836	F	
42	344459977542134	5686270	M	
43	33347277946935	5686271	M	
44	6672769957823	5574844	F	
45	55741363563825	5602557	F	
46	12217876371278	5686275	M	
47	524654338535	5647550	F	
48	897998656873733	5642546	F	
49	388579184536169	5701581	F	
50	57774856817778	5753226	M	
51	697121536733778	5594152	F	
52	61488471593129	5681294	F	
53	321484469257929	5574366	F	
54	92354424593638	5579997	F	
55	273582875734917	5701444	F	
56	11334367453584	5654332	F	
57	912364259987	5560383	F	

34	191395636276764	5657142	F	22 JAR
35	9356651735151	5658036	M	?
36	21833136218142	5653210	F	41 JAR
37	31277536312512	5561250	F	62 JAR
38	5914418184422	5580234	F	83 JAR
39	297793143163734	5602465	M	74 JAR
40	743733582334942	5602969	F	38 CEN
41	17385944962	5624836	F	27 JAR
42	344459977542134	5686270	M	24 JAR
43	33347277946935	5686271	M	77 JAR
44	6672769957823	5574844	F	21 JAR
45	55741363563825	5602557	F	24 JAR
46	12217876371278	5686275	M	18 JAR
47	524654338535	5647550	F	38 JAR
48	897998656873733	5642546	F	58 JAR
49	388579184536169	5701581	F	6 JAR
50	57774856817778	5753226	M	18 JAR

How?

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

Can Also Use Filter to Identify Missing Data

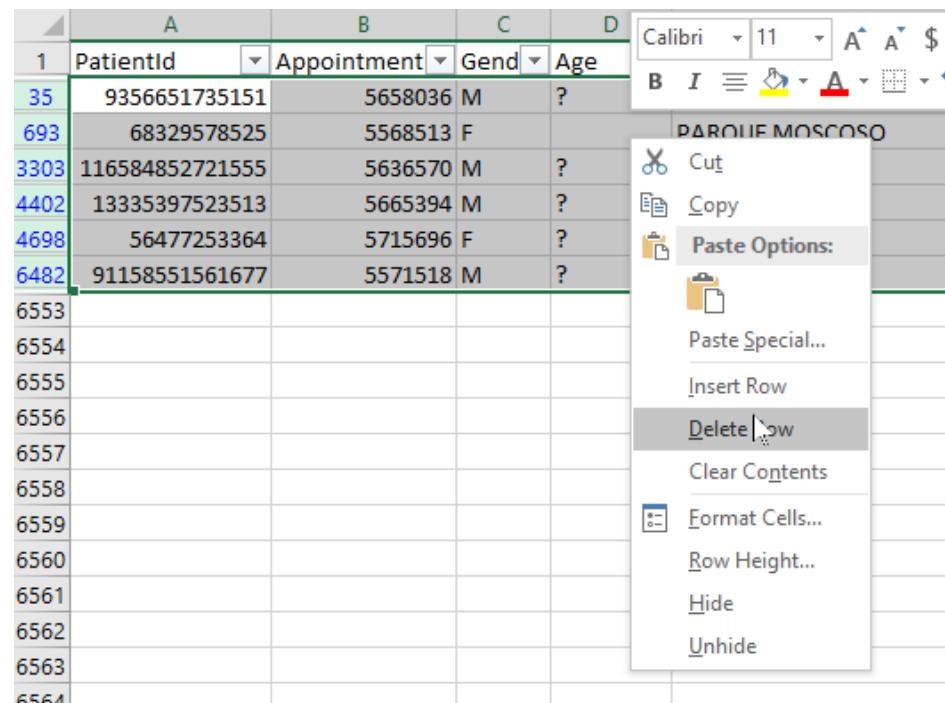
The figure consists of three side-by-side screenshots of Microsoft Excel spreadsheets. The first two screenshots show the 'Filter' dialog box open over a data range. The first dialog shows the 'Age' column being filtered with 'Select All' checked. The second dialog shows the 'Age' column being filtered with 'Blanks' checked. The third screenshot shows the final result of the filtering applied across all columns.

A	B	C	D	E	
1	PatientId	Appointment	Gend	Age	Neighbourhood
2	29872499	Z ↓ Sort Smallest to Largest	JF		
3	558997776	Z ↓ Sort Largest to Smallest	JF		
4	2978363	Sort by Color	JF		
5	6326444	Clear Filter From "Age"	JF		
6	848511136	Filter by Color	JF		
7	5794444	Number Filters	JF		
8	21215964	Search	JF		
9	16723534	(Select All)	JF		
10	995196886	0	JF		
11	6736855	1	JF		
12	62661375	2	JF		
13	994495	3	JF		
14	83391743	4	JF		
15	925953	5	JF		
16	86497	6	JF		
17	616161776	7	JF		
18	68182861	8	JF		
19	71994897	9	JF		
20	176136196	10	JF		
21	68838234	11	JF		
22	92282526182277	12	JF		
		13	JF		
		14	JF		
		15	JF		
		16	JF		
		17	JF		
		18	JF		
		19	JF		
		20	JF		
		21	JF		
		22	JF		

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



A screenshot of a Microsoft Excel spreadsheet. The data is organized into columns A through D. Column A contains patient IDs, column B contains appointment numbers, column C contains gender, and column D contains age. Row 6546 is highlighted with a red border. A context menu is open over this row, with the 'Delete Row' option highlighted in grey. To the right of the table, a separate section shows rows 6544, 6545, 6546, 6547, 6548, and 6549. Row 6546 is also highlighted with a red border. An orange callout bubble points to this row with the text 'Last row number is 6546'.

1	PatientId	Appointment	Gend	Age
35	9356651735151	5658036	M	?
693	68329578525	5568513	F	
3303	116584852721555	5636570	M	?
4402	13335397523513	5665394	M	?
4698	56477253364	5715696	F	?
6482	91158551561677	5571518	M	?
6553				
6554				
6555				
6556				
6557				
6558				
6559				
6560				
6561				
6562				
6563				
6564				

6544	5965757389437	5771390	F	56 JARDIM CAMBURI
6545	7825443833792	5767544	F	60 JARDIM CAMBURI
6546	598674846832672	5770793	M	44 JARDIM CAMBURI
6547				
6548				
6549				

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

	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	55899776694438	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

The screenshot shows the Excel interface with a data table and two filter dialog boxes.

Left Panel: A context menu is open over the 'WaitingDays' column header, showing options like 'Sort Smallest to Largest', 'Sort Largest to Smallest', and 'Number Filters'. The 'Number Filters' option is selected, opening the 'Custom AutoFilter' dialog.

Custom AutoFilter Dialog: This dialog is titled 'Custom AutoFilter' and 'Show rows where: WaitingDays'. It contains a dropdown menu set to 'is less than' with the value '0'. Below it is a radio button group for 'And' and 'Or', with 'And' selected. There are also dropdown menus for 'Between...' and 'Top 10...'. At the bottom are 'OK' and 'Cancel' buttons.

Data Table: The main table has columns labeled from A to P. Row 1 contains column headers. Rows 2 and 3 show patient data. Row 4 is a blank row labeled '6547'. Rows 5 and 6 are redacted. Row 7 is highlighted with a red border and contains the number '6544'. Row 8 is a blank row labeled '6545'. Row 9 is a blank row labeled '6546'. An orange callout bubble points to the number '6544' in row 7, with the text 'Last row number is 6544'.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	PatientId	Appointment	Gend	Age	Neighbourhood	FamilyAllowanc	Hypertensic	Diabet	Alcoholis	Handic	SMS_receive	No-shc	AppointmentDa	ScheduleDa	ScheduleTim	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
6542	7825442922792	5767544	F	60	JARDIM CAMBURI	0	0	0	0	0	1	No	2016-06-07	2016-06-02	18:22:06	5
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																

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

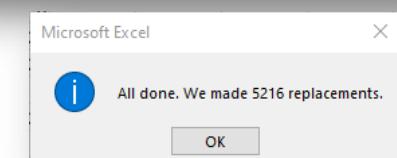
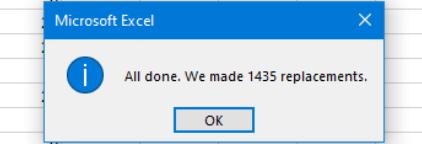
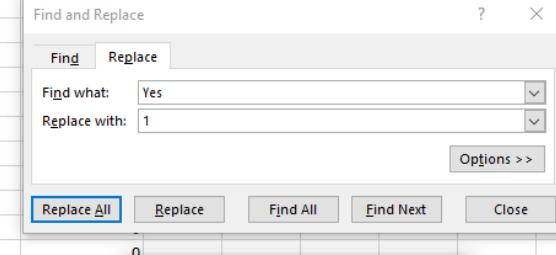
FIND AND REPLACE TEXTS

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	Yes	2016-05-20
0	0	No	2016-05-20
0	0	No	2016-05-05

L	M	N	O	P	Q	R	S	T	U
No-show	AppointmentDate	ScheduleDate	ScheduleTime	WaitingDays					
No	2016-04-29	2016-04-29	18:38:08	0					
No	2016-04-29	2016-04-29	16:08:27						
No	2016-05-12	2016-05-06	17:10:31						
No	2016-05-05	2016-05-05	09:14:10						
No	2016-05-06	2016-04-08	12:56:06						
No	2016-05-13	2016-04-15	08:37:18						
No	2016-05-20	2016-04-25	15:58:50						
No	2016-05-06	2016-05-06	09:38:36						
No	2016-05-13	2016-05-13	09:47:09						
No	2016-05-06	2016-05-06	10:02:59						
No	2016-05-13	2016-05-13	10:41:09						
No	2016-05-20	2016-05-20	10:54:26						
No	2016-05-06	2016-04-08	18:18:01						
No	2016-05-13	2016-04-15	10:17:19						
No	2016-05-06	2016-05-06	08:34:17						
No	2016-05-13	2016-04-15	11:00:48						
1	2016-05-20	2016-05-11	17:27:44						
No	2016-05-20	2016-05-20	11:29:35						



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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	0	2016-05-13	2016-04-15	11:00:48	28
18	68182861542773	5687942	F	77	JARDIM DA PENHA	0	1	0	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	0	2016-05-20	2016-05-20	11:29:35	0
6537	17327699252721	5764674	F	62	JARDIM CAMBURI	0	1	0	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	0	2016-06-06	2016-06-02	10:38:33	4
6539	588313356244741	5763362	F	60	JARDIM CAMBURI	0	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	0	2016-06-06	2016-06-06	17:42:01	0
6541	19314223565	5767006	F	55	JARDIM CAMBURI	0	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	0	2016-06-07	2016-06-03	12:45:35	4
6543	7825443833792	5767544	F	60	JARDIM CAMBURI	0	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	0	2016-06-07	2016-06-03	10:57:50	4
6545																

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	0 Child	
Senior	13 Adolescence	
Senior	19 YoungAdult	
YoungAdult	46 MiddleAge	
Senior	60 Senior	

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	2nd column
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		

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		5 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		38 Senior		
7	5587385	F	57	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-13	2016-04-15	08:37:18		38 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		35 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		38 Senior		
15	5588626	M	62	JARDIM DA PENHA	0	0	0	0	0	0	0	2016-05-13	2016-04-15	10:17:19		38 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		38 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		0 MiddleAge		
6542	5771390	F	56	JARDIM CAMBURI	0	0	0	0	0	1	0	2016-06-07	2016-06-03	12:45:35		0 MiddleAge		
6543	5767544	F	60	JARDIM CAMBURI	0	0	0	0	0	1	0	2016-06-07	2016-06-02	18:32:06		0 Senior		
6544	5770793	M	44	JARDIM CAMBURI	0	0	0	0	0	1	0	2016-06-07	2016-06-03	10:57:50		0 YoungAdult		
6545																		
6546																		
6547																		

FIND OUTLIERS

Outliers in WaitingDays

Calculate Quartiles

$Q1 = \text{QUARTILE}(P1:P6544, 1)$

$Q3 = \text{QUARTILE}(P1:P6544, 3)$

Calculate Upper and Lower Boundaries

$IQR = Q3 - Q1$

$L \text{ BOUND} = Q1 - IQR * 1.5$

$U \text{ BOUND} = Q3 + IQR * 1.5$

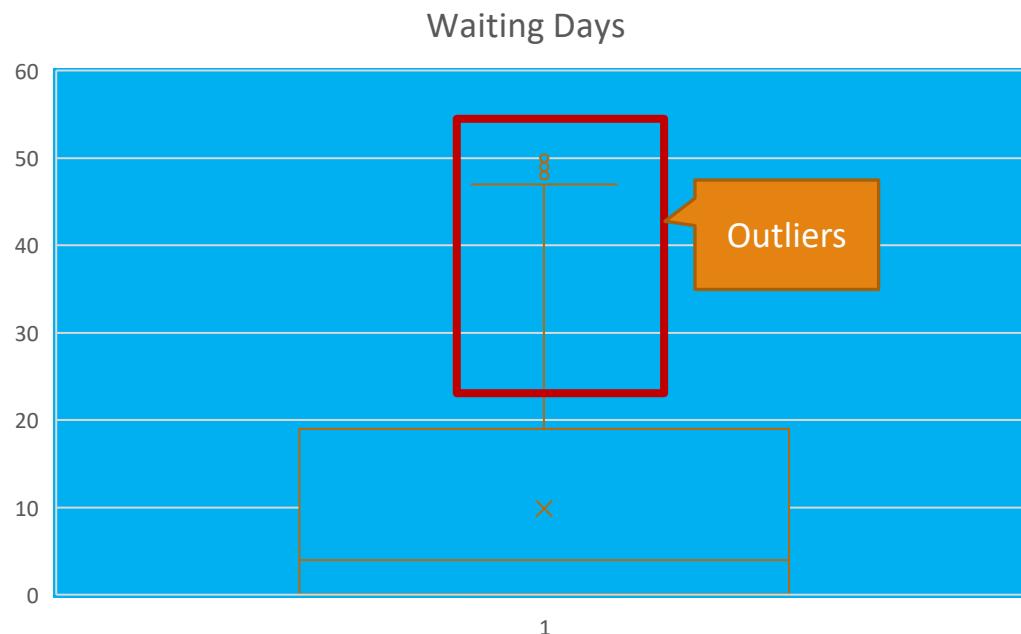
$\text{Outlier} = \text{WaitingDays} < LB \text{ or } \text{WaitingDays} > UB$

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	Q1	0		
28	Q3	20		
0		YoungAdult	IQR	
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

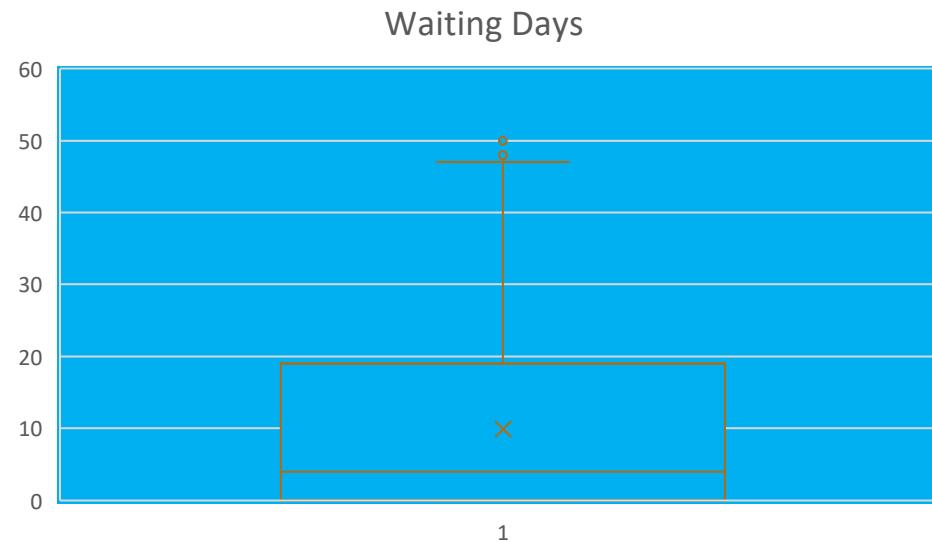
There are 96 outliers
(revealed with filtering TRUE)

The screenshot shows the 'Outlier' filter dialog open over a data table. The dialog has three filter options: 'Sort Smallest to Largest', 'Sort Largest to Smallest', and 'Sort by Color'. Below these are 'Clear Filter From "Outlier"', 'Filter by Color', and 'Number Filters'. Under 'Number Filters', the 'TRUE' checkbox is selected. The main table to the right shows a column labeled 'Outlier' with many entries set to 'TRUE', indicating they are outliers. A red box highlights this column. At the bottom of the dialog are 'OK' and 'Cancel' buttons. To the right of the table, a status bar says 'Ready 96 of 6543 records found'. An orange callout box points to the 'Outlier' column with the text '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																
6451																

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