

Transcript | Hacks for Cleaning Messy Data (in Minutes)

1

00:00:11.870 --> 00:00:25.989

Joel Smith: Hello! Hello, everybody, and welcome to today's Masterclass event, Hacks for Cleaning Messy Data in Minutes. My name is Joel Smith, and I'm the Events Manager here at Apolitical.

2

00:00:26.030 --> 00:00:44.410

Joel Smith: Today, we're looking at how public servants can clean messy spreadsheets quickly, safely, and with confidence. We'll hear how to spot the issues that block analysis, from duplicates and inconsistent categories to missing fields and mixed date formats, and how to fix them without losing the information that matters.

3

00:00:44.430 --> 00:00:53.919

Joel Smith: For those of you who may be joining us for the first time, first of all, welcome. Apolitical is the world's largest online network for government. For more than

4

00:00:53.920 --> 00:01:13.530

Joel Smith: half a million public servants and policymakers across 170 countries. Apolitical is a strategic partner to governments. We help them prepare their workforces for the future. We achieve this by upskilling through short online courses and peer communities, as well as by building innovative learning tools designed for and with governments.

5

00:01:13.670 --> 00:01:29.349

Joel Smith: So, it would be great to hear from all of you too, so please do post in the chat where you're joining us from and what organization you work for. I can already see a couple in there. I'm just gonna get through a couple of quick bits of housekeeping before we get into today's session.

6

00:01:29.350 --> 00:01:36.470

Joel Smith: Firstly, I'll mention that the event is recorded for on-demand viewing on the Apolitical website, so please don't record it yourself.

7

00:01:36.510 --> 00:01:42.249

Joel Smith: We'll also be using the recording to generate events insights, including with AI-powered software.

8

00:01:42.330 --> 00:01:50.640

Joel Smith: As this session is running on a Zoom webinar, we won't collect your video or audio data, but personal data you share in the chat might be captured.

9

00:01:50.680 --> 00:02:05.110

Joel Smith: Contact details for any privacy concerns are in the Apolitical homepage footer, and closed captions are available in your own language and can be toggled on or off in your screen settings. My colleague, I believe, has just posted the instructions in the chat.

10

00:02:05.160 --> 00:02:20.410

Joel Smith: And secondly, before we get started, I would love to introduce you to the Data Skills Network on Apolitical. It's a space to ask the questions we don't always get time for at work, or at these events, such as how to format data imported from PDFs.

11

00:02:20.440 --> 00:02:34.580

Joel Smith: This community is a great place to connect with and learn from your peers in the public sector on all things data. My colleague will also post a link to the community in the chat below. Okay, I'm just going to quickly jump into the chat and see who we have.

12

00:02:34.680 --> 00:02:49.710

Joel Smith: Hello, first comment from Nova Scotia, welcome. We've got London, we've got Sheffield representing, we have South Carolina in the US, more from the US, Canada, Ontario, Jersey...

13

00:02:49.810 --> 00:03:11.940

Joel Smith: York, Canada, we have someone joining from Aberdeen, Leeds. It's lovely to see you all. These comments are coming thick and fast, but a nice, strong representation across, the UK and Canada. Some folks from the US, Cape Town, South Africa. Welcome, welcome. It's lovely to have you all here.

14

00:03:13.470 --> 00:03:25.880

Joel Smith: Perfect! So, we're gonna get into, the main event of today, introducing our speaker. We're very fortunate to have Alan Yarish joining us today.

15

00:03:25.880 --> 00:03:35.079

Joel Smith: Alan Yarish is a finance professional with more than 17 years of experience across investment fund marketing, portfolio management, and financial analysis.

16

00:03:35.080 --> 00:03:58.750

Joel Smith: He is currently a partner at Lancea Partners, having previously worked as a portfolio manager at Channel Capital Advisors. Alongside his extensive industry experience, he is an enthusiastic and highly skilled Excel user, with a particular interest in harnessing the seamlessly... the seemingly limitless power of spreadsheets to analyze data, solve problems, and improve decision making.

17

00:03:58.900 --> 00:04:05.349

Joel Smith: You can read the full bio by visiting the resources tab at the bottom of your screen and heading to our speaker section.

18

00:04:05.520 --> 00:04:24.110

Joel Smith: But, without further ado, Alan, it's an absolute pleasure to have you with us. I'm going to hand over to you now for the next 15 minutes or so, and then we're going to have a few minutes reserved for Q&A. So, attendees, please do submit your questions for Alan in the Q&A tab as he's speaking. But, Allan, the floor is yours.

19

00:04:24.790 --> 00:04:33.580

Allan Yarish: Thank you very much, Joel. Let me just start by... I'll share the spreadsheet that we're going to be using today.

20

00:04:38.710 --> 00:04:40.340

Allan Yarish: And,

21

00:04:40.570 --> 00:04:55.579

Allan Yarish: So this is just some demo data, and what we're trying to achieve here is to show you a few things about how to display data and how to fix some data problems. And Joel and his colleagues suggested we use this, which is some Icelandic

22

00:04:55.880 --> 00:05:09.170

Allan Yarish: weather reporting, it's just the number of sunshine hours on a given day. I thought... I know where Reykjavik is, but I didn't know any of the other names, so just for a moment's interest, there's what it looks like on a map.

23

00:05:09.300 --> 00:05:16.220

Allan Yarish: of Iceland. So, you can see it's basically different, areas around the perimeter of Iceland.

24

00:05:16.460 --> 00:05:20.210

Allan Yarish: So, a couple of very simple things.

25

00:05:20.270 --> 00:05:30.939

Allan Yarish: Very, very basic things. One of the things... this is a small table, but you may well have a table that has, like, 500 rows in it and, you know, 25 columns.

26

00:05:30.960 --> 00:05:46.149

Allan Yarish: And if you scroll down like that, you can see it disappears. A very simple thing that you want to know if you're not already doing it is how to, freeze the top right columns and rows.

27

00:05:46.270 --> 00:05:52.040

Allan Yarish: The way you do that is you click on View at the top here. There's something called Freeze Panes.

28

00:05:52.220 --> 00:05:59.410

Allan Yarish: And it'll freeze the top right and top left columns. So you can see if we scroll down.

29

00:05:59.650 --> 00:06:04.520

Allan Yarish: The, titles at the top stay, and if we scroll to the right.

30

00:06:05.290 --> 00:06:07.870

Allan Yarish: The first column stays in place.

31

00:06:07.980 --> 00:06:10.600

Allan Yarish: So, just a very simple thing.

32

00:06:12.730 --> 00:06:21.990

Allan Yarish: you'll see that this data has a few issues with it. What we've got is the name of the weather station, the region that it's in within Iceland.

33

00:06:22.110 --> 00:06:30.639

Allan Yarish: the date that it was taken, and the number of sunshine hours, as well as some notes and comments. And a couple of things...

34

00:06:30.760 --> 00:06:36.470

Allan Yarish: Here, just to highlight problems that we're going to address, you can see the date formats.

35

00:06:37.430 --> 00:06:48.000

Allan Yarish: have not all been consistently entered. So, most of them are using this format, where you have two digits for the day, a slash 2 digits for the month, and four digits for the year.

36

00:06:48.010 --> 00:06:57.720

Allan Yarish: But if you look at this, row number 7, you can see they're using decimal points instead of slashes. And also, you'll see that in some cases.

37

00:06:57.800 --> 00:07:06.850

Allan Yarish: they are using American date formats rather than European date formats. That's a very consistent problem.

38

00:07:07.020 --> 00:07:18.500

Allan Yarish: The second thing is, not everybody has filled in the data. There are two rows that are blank here. So, we'll deal with that in a second. So... and some of the data appears to have been entered twice.

39

00:07:18.830 --> 00:07:23.219

Allan Yarish: So the first thing we're going to do is learn how to remove duplicate data.

40

00:07:23.650 --> 00:07:29.670

Allan Yarish: And... This is done very simply. First thing you do is highlight the table.

41

00:07:29.960 --> 00:07:37.499

Allan Yarish: that you want to, eliminate duplicates... duplicates on. You click on the Data tab up at the top.

42

00:07:38.220 --> 00:07:47.559

Allan Yarish: And, there's an entry called Remove Duplicates here. Now, one thing to notice is you may, depending on

43

00:07:47.790 --> 00:07:52.820

Allan Yarish: how you set up Excel. You may have This view only.

44

00:07:53.280 --> 00:07:58.960

Allan Yarish: You have to click on data, and all of these options come down below you.

45

00:07:59.270 --> 00:08:05.479

Allan Yarish: You control that by either Show Tabs Only or Always Show Ribbon, which is typically how I do it.

46

00:08:05.700 --> 00:08:09.189

Allan Yarish: However you do it, you click on Remove Duplicates here.

47

00:08:10.790 --> 00:08:21.709

Allan Yarish: And this list of all the columns... now, first of all, an important thing to note, and this is something that's common to a number of data table operations, you'll see it says, my data has headers.

48

00:08:22.100 --> 00:08:35.860

Allan Yarish: If you... if that is checked, which it should be for most tables that you're going to work with, it leaves the top row as the headers and doesn't apply any of the selection criteria to the top row, only to the data rows below that.

49

00:08:36.120 --> 00:08:41.409

Allan Yarish: But it's showing our four data rows, station, date, sunshine hours, and reported by.

50

00:08:41.520 --> 00:08:53.600

Allan Yarish: If we click all of those, it's going to eliminate any second row which is identical to another row with the same information in all four of those columns. So let's click on that.

51

00:08:54.280 --> 00:09:00.710

Allan Yarish: And you get this message, 3 duplicate values found and removed. 11 unique values remain.

52

00:09:01.280 --> 00:09:03.949

Allan Yarish: Alright, I'll... but I want to show you... so...

53

00:09:04.940 --> 00:09:11.609

Allan Yarish: I'm going to reverse that now so that I can show you some other options. So we click the undo button up in the top left.

54

00:09:12.790 --> 00:09:17.789

Allan Yarish: We can do, remove duplicates, and we could just decide

55

00:09:18.050 --> 00:09:26.310

Allan Yarish: We don't care about the date, we don't care about the sunshine hours, we don't care about who it was reported by, we just want to remove any duplicates where it's just the same

56

00:09:26.660 --> 00:09:29.029

Allan Yarish: Station name in the first column.

57

00:09:29.290 --> 00:09:32.650

Allan Yarish: And that gives us 4 duplicate values found and removed.

58

00:09:33.000 --> 00:09:39.579

Allan Yarish: And you could do this in a variety of different ways. We could also do it by...

59

00:09:40.420 --> 00:09:46.149

Allan Yarish: the reporting... Station, so...

60

00:09:46.250 --> 00:09:48.469

Allan Yarish: If we just click on Reported By.

61

00:09:49.350 --> 00:09:57.480

Allan Yarish: it gives us 8 duplicate values were found. Generally, you want Oops,

62

00:09:59.420 --> 00:10:08.420

Allan Yarish: Generally, you want to do it on all of the columns, or certainly you would want, possibly, to do it on all the ones with the same station and the same date.

63

00:10:08.480 --> 00:10:22.569

Allan Yarish: perhaps you'd be willing to tolerate differences. But that's a simple way to deal with messy data, and you'll find this if your data source is a bunch of different people inputting data, or...

64

00:10:23.090 --> 00:10:31.809

Allan Yarish: importing it at different times of the day, you may well end up with a number of duplicates. So that's a very simple way to remove duplicates.

65

00:10:32.610 --> 00:10:36.930

Allan Yarish: Another problem that you'll see is...

66

00:10:37.030 --> 00:10:46.940

Allan Yarish: People... this table, we've shown... the names are spelled in different ways, and you can see there's some blanks, or there are some

67

00:10:47.170 --> 00:10:57.249

Allan Yarish: spaces between the names, and you'd like to standardize that. So there are a bunch of different functions, and we're going to show you

68

00:10:57.560 --> 00:11:00.429

Allan Yarish: how to use the function, so...

69

00:11:01.070 --> 00:11:04.979

Allan Yarish: for those who are not familiar with functions in Excel, you start by

70

00:11:05.100 --> 00:11:12.690

Allan Yarish: entering the equal sign, and then you... you need to know what the function name is. In this case, we're going to use the function called trim.

71

00:11:13.520 --> 00:11:17.860

Allan Yarish: And then you highlight that cell, the source cell.

72

00:11:18.180 --> 00:11:25.439

Allan Yarish: And what trim does is it strips out any leading or trailing zero blanks in the name.

73

00:11:25.690 --> 00:11:33.549

Allan Yarish: And so let's just... We've done it on the first cell, we're now going to copy it, so... copy...

74

00:11:35.790 --> 00:11:36.840

Allan Yarish: And paste.

75

00:11:37.350 --> 00:11:46.639

Allan Yarish: So you can see that what Trim has done in this second line, where there's a bunch of blanks in front of the capital letter version of Reykjavik, it's,

76

00:11:47.020 --> 00:11:49.250

Allan Yarish: It's eliminated those blanks.

77

00:11:49.660 --> 00:11:54.540

Allan Yarish: But you'll still notice there's a mix of uppercase and lowercase here.

78

00:11:55.040 --> 00:11:58.709

Allan Yarish: And there are a variety of choices, 3 different functions.

79

00:11:58.910 --> 00:12:02.379

Allan Yarish: One's called Proper, and what Proper does

80

00:12:02.580 --> 00:12:07.449

Allan Yarish: Is it converts this text to having,

81

00:12:07.890 --> 00:12:11.549

Allan Yarish: A capitalized first letter and everything else in lowercase.

82

00:12:11.910 --> 00:12:15.220

Allan Yarish: And... Again, if we copy that down.

83

00:12:18.270 --> 00:12:30.979

Allan Yarish: you can see that, so in this second line, the capital letter version of Reykjavik becomes Reykjavik with a capital first letter and everything else in lowercase. If we want everything in uppercase, we can do that with

84

00:12:31.870 --> 00:12:37.219

Allan Yarish: upper... And if we want everything in lowercase, we can do that with lower.

85

00:12:40.510 --> 00:12:42.609

Allan Yarish: And then we can copy that down.

86

00:12:44.730 --> 00:12:53.230

Allan Yarish: So, generally, you're not going to want to do all three of these. I think in most purposes, the proper is a good way to standardize the data.

87

00:12:54.750 --> 00:13:03.690

Allan Yarish: So that's probably what you would use, but you may want uppercase or lowercase, depending on the situation. Another thing you might want to do

88

00:13:04.170 --> 00:13:08.920

Allan Yarish: If you're sending out a table for other people, to...

89

00:13:09.430 --> 00:13:22.229

Allan Yarish: enter the data is use something called data validation. So we have over on... over here, there's a list of the names and how we want them spelled, and because this is in Icelandic, you see there's a number of accents and

90

00:13:22.640 --> 00:13:41.509

Allan Yarish: various other characters, you might want to insist that people always do it consistently. You can see some people are using the accents over here, and some people aren't. One way to do that is you highlight the cells in which you want the data entered, you click on the data tab again, and go to something called Data Validation.

91

00:13:42.210 --> 00:13:45.470

Allan Yarish: What data validation is going to do is

92

00:13:45.950 --> 00:13:50.460

Allan Yarish: If you were going to say that the names have to be chosen from a list.

93

00:13:50.750 --> 00:13:55.049

Allan Yarish: And the list that we're going to highlight is this table over here.

94

00:13:57.100 --> 00:13:58.989

Allan Yarish: And so it says.

95

00:14:00.860 --> 00:14:06.879

Allan Yarish: If we're going to enter a name over there, it has to be a name that appears on this source list under the controlled list status.

96

00:14:07.170 --> 00:14:07.960

Allan Yarish: tab.

97

00:14:08.160 --> 00:14:10.900

Allan Yarish: So, let's just try to... if we...

98

00:14:11.060 --> 00:14:19.059

Allan Yarish: try to put this in. We're now going to get an error that says this value doesn't matter... match the data validation restrictions for this cell.

99

00:14:20.430 --> 00:14:27.969

Allan Yarish: But you notice there's a little arrow on the side here, and so we can... if we click the arrow, it gives us the choice of

100

00:14:28.410 --> 00:14:35.520

Allan Yarish: everything that we... that's on that controlled list, and that allows us to do it. So...

101

00:14:36.060 --> 00:14:50.790

Allan Yarish: That's a very simple way. It's important if you're sending out worksheets for other people to do. You want... you'll find data validation is a useful way to get consistency in the data and avoid having to clean it up.

102

00:14:51.500 --> 00:14:52.830

Allan Yarish: After the mess.

103

00:14:53.030 --> 00:15:01.369

Allan Yarish: The last thing I'll talk about, On this is, date formats, so...

104

00:15:01.730 --> 00:15:16.610

Allan Yarish: We have a number of people from Europe and from North America. If you're used to dealing with spreadsheets across different continents, you'll know that American date formats are somewhat different from Excel date formats.

105

00:15:16.960 --> 00:15:22.110

Allan Yarish: I want to start with something very basic, which is the way Excel records dates

106

00:15:22.670 --> 00:15:27.799

Allan Yarish: You can see this if we... we just write a little formula to copy a date here.

107

00:15:28.020 --> 00:15:32.270

Allan Yarish: And... We'll format that as...

108

00:15:35.520 --> 00:15:36.760

Allan Yarish: a number.

109

00:15:41.080 --> 00:15:41.850

Allan Yarish: -Oh.

110

00:15:49.950 --> 00:15:52.520

Allan Yarish: Why isn't that non-formatting?

111

00:16:00.380 --> 00:16:03.220

Allan Yarish: Okay, gotta do something else here.

112

00:16:15.360 --> 00:16:19.660

Allan Yarish: Okay, there's a good example. So... Excel...

113

00:16:19.860 --> 00:16:28.119

Allan Yarish: displays dates in a variety of different ways. It has a date, which is a, it uses a serial number.

114

00:16:28.950 --> 00:16:34.719

Allan Yarish: The serial number is based on a convention that January 1st, 1990,

115

00:16:34.840 --> 00:16:41.080

Allan Yarish: is day number one, and January 2nd is 1900, sorry.

116

00:16:41.200 --> 00:16:52.080

Allan Yarish: is day number 2, and so on. And so, now we're in 2026, and it's... it's, day number 46,183 since the start of the 20th century.

117

00:16:52.410 --> 00:16:59.340

Allan Yarish: So underlying everything that happens with, dates is this serial number, but...

118

00:16:59.670 --> 00:17:04.899

Allan Yarish: Dates are displayed. In Europe, they're displayed. The first two digits are

119

00:17:04.900 --> 00:17:22.739

Allan Yarish: the day, the second two digits are the month, and the last two or four digits are the year. In America, it's the first two digits are the month, the second two digits are the day, the second two digits are the day, and the last two or four are the year. So that creates some confusion sometimes.

120

00:17:23.670 --> 00:17:28.089

Allan Yarish: What we've done here, and you can see in this table, we've got

121

00:17:28.740 --> 00:17:36.000

Allan Yarish: several dates, like, this would be, on a European date format, the 3rd of June, the 6th of March.

122

00:17:36.310 --> 00:17:38.719

Allan Yarish: the 15th of June.

123

00:17:38.930 --> 00:17:49.310

Allan Yarish: But here, in row number 7, clearly, this is an American date format, because this has to be the 15th of June. There's no such thing, there's no 15th month.

124

00:17:49.490 --> 00:17:57.190

Allan Yarish: So... we've got to sort that out. A very simple way to do that is just

125

00:17:57.540 --> 00:18:00.339

Allan Yarish: If you write a formula

126

00:18:01.680 --> 00:18:16.529

Allan Yarish: where you add one to a date. It'll come up as a number if the source date is a date, but it'll give you this value error if there's something inconsistent in these source dates. So in these two

127

00:18:16.740 --> 00:18:36.109

Allan Yarish: As I mentioned, this is a spreadsheet using European date conventions. The row number 7 is clearly written with an American date convention, and so it's not recognized. And

row number 8, the digits are right, but the separators are the decimal points rather than the slashes, so it's not recognizing that either.

128

00:18:36.450 --> 00:18:41.109

Allan Yarish: And similarly, down here, we have an American date format.

129

00:18:41.290 --> 00:18:47.859

Allan Yarish: so, a very simple... we're gonna deal with this in a number of steps.

130

00:18:48.200 --> 00:18:58.560

Allan Yarish: The first thing here is we're going to use something called the substitute function, and I've written this out already, but basically, substitute is a function which allows you to

131

00:18:58.860 --> 00:19:03.679

Allan Yarish: substitute The slashes for the decimal points.

132

00:19:04.210 --> 00:19:09.910

Allan Yarish: And that's going to... Changed one cell here in particular.

133

00:19:10.180 --> 00:19:18.670

Allan Yarish: So, 7.06.2026 now becomes 07 slash 06 slash 2026.

134

00:19:19.310 --> 00:19:20.870

Allan Yarish: And...

135

00:19:21.750 --> 00:19:37.930

Allan Yarish: that allows us to start treating these as dates. The second thing we're going to do is write a little formula that allows us to figure out whether this is a recognizable date format, whether it's consistent with European dating conventions. And we're going to use two formulas there.

136

00:19:38.920 --> 00:19:42.130

Allan Yarish: The first is date value.

137

00:19:42.400 --> 00:19:45.360

Allan Yarish: And date value just says,

138

00:19:45.860 --> 00:19:50.890

Allan Yarish: Gives you the, the number, associated with that date.

139

00:19:51.130 --> 00:19:53.560

Allan Yarish: In the Excel serial codes.

140

00:19:53.920 --> 00:20:05.770

Allan Yarish: And then we're gonna wrap around that another function which says isNUMBER, and that's just a... what's called a logical variable that is either true or false. So, essentially, what we're doing is

141

00:20:05.970 --> 00:20:14.560

Allan Yarish: calculating whether the date value generates a serial code and is recognized as a date value in Excel.

142

00:20:14.670 --> 00:20:19.999

Allan Yarish: You can see that we've got two here that have come back false, and these are the ones where the second

143

00:20:20.230 --> 00:20:28.159

Allan Yarish: value in the digits is clearly not a month, because it's a number that's higher than 12. So this one's false.

144

00:20:29.080 --> 00:20:31.840

Allan Yarish: This line down here is false as well.

145

00:20:32.600 --> 00:20:46.909

Allan Yarish: So, and I'm probably getting a little too complicated for anybody who's just a beginner in Excel. We've come up with formulas here. They are a little messy. They are... use a combination... and I've listed down below the various

146

00:20:47.010 --> 00:20:52.749

Allan Yarish: Now, so far we've used substitute, is number, and, and date value.

147

00:20:53.060 --> 00:20:56.349

Allan Yarish: Over here. Now we're going to be using the...

148

00:20:57.620 --> 00:21:07.079

Allan Yarish: these functions, value, left, mid, right, and find. And value simply just gives you the numerical value of something. Left.

149

00:21:07.250 --> 00:21:12.239

Allan Yarish: strips out the left, the, the...

150

00:21:12.580 --> 00:21:16.330

Allan Yarish: strings in the left-hand side of a string.

151

00:21:16.650 --> 00:21:21.669

Allan Yarish: Have to use a number, you... how many characters you want it to take out of that string.

152

00:21:21.890 --> 00:21:27.290

Allan Yarish: So we use FIND to look for the first occurrence of the slash

153

00:21:27.510 --> 00:21:38.190

Allan Yarish: In the... in the source here. Anyway, all that to say, it strips it out into the... we've taken 03 slash 06 slash 2026.

154

00:21:38.360 --> 00:21:43.720

Allan Yarish: And with these formulas, have split them up into

155

00:21:44.770 --> 00:21:52.370

Allan Yarish: a simple 3, a simple 6, and a 2026. We then use... the date variable.

156

00:21:53.070 --> 00:21:56.609

Allan Yarish: to rebuild these as dates, but...

157

00:21:56.900 --> 00:22:03.260

Allan Yarish: I built in an if function, and the if function checks whether

158

00:22:05.400 --> 00:22:08.440

Allan Yarish: If it's already recognized as a date.

159

00:22:08.680 --> 00:22:20.330

Allan Yarish: then we put it together using the European date format. If it's not recognized as a date, we swap the order of the day and the month, and it comes out. So, in this line.

160

00:22:20.950 --> 00:22:35.470

Allan Yarish: Where it's, you're not recognized as a date because the second digit is greater than 12, it says, okay, well, it must be the 15th of June rather than the 6th of a month that doesn't exist. And so that sorts it all out.

161

00:22:37.180 --> 00:22:47.820

Allan Yarish: I'm happy to go... when I was putting this together, I realized that these formulas are somewhat complicated, and unfortunately, you probably have a little bit of

162

00:22:48.130 --> 00:22:49.450

Allan Yarish: difficulty.

163

00:22:49.760 --> 00:22:58.890

Allan Yarish: recognizing them, because I'm not sure that we managed to zoom out the main body of the spreadsheet, but the formulas up top here are a little bit small.

164

00:23:00.370 --> 00:23:10.689

Allan Yarish: But those are the main things we wanted to cover today. I don't know, Joel, if we... we still have some time left. I can go into these formulas in a little more detail if you want me to.

165

00:23:11.060 --> 00:23:26.780

Joel Smith: We've got a few questions. I'm aware we've only got 7 minutes, but, if it's okay with you, if you could send us through those formulas, and then we could share those in the

follow-up email with everyone, so they can, they can see them in more detail and have a play themselves.

166

00:23:27.430 --> 00:23:29.999

Allan Yarish: Okay, well, I could copy them down here.

167

00:23:31.920 --> 00:23:35.260

Joel Smith: Oh, perfect, yeah, that would be good, and then everyone should be able.

168

00:23:35.260 --> 00:23:37.930

Allan Yarish: Would it be a little bit more visible.

169

00:23:38.570 --> 00:23:40.879

Allan Yarish: And we can walk through what they do.

170

00:23:46.000 --> 00:23:47.270

Allan Yarish: Oh, hang on.

171

00:23:56.530 --> 00:24:12.099

Joel Smith: I think, yeah, let's maybe, if you send us those, we'll share these with the, with the wider audience, and we can, yeah, we can, we can share those in that way. And if it's okay, yeah, we do have a couple of questions.

172

00:24:12.100 --> 00:24:17.149

Joel Smith: Which I'd just like to try and blast through before, before we wrap.

173

00:24:17.470 --> 00:24:26.869

Joel Smith: So, firstly, thank you so much for that. That was so, so brilliant, so detailed. I've definitely learned a lot, and I've been taking a lot of notes.

174

00:24:26.970 --> 00:24:34.760

Joel Smith: Just before we get into the questions, I just want to let everyone know that we have developed a follow-on resource from today's topic.

175

00:24:34.760 --> 00:24:48.069

Joel Smith: which is a short lesson from one of our courses on what data confidence means, and how to start building data confidence in your day-to-day work. So please do, I've posted in the data skills community, my colleague is just showing that in the chat.

176

00:24:48.070 --> 00:24:54.649

Joel Smith: So please do, comment on there, and we will share that lesson with you, and we will share,

177

00:24:54.690 --> 00:25:09.169

Joel Smith: these formulas that Allan has run through in the follow-up email that comes over tomorrow as well. So, swiftly on to the Q&A, so just a couple of questions we'll see if we can blast through before we end.

178

00:25:09.420 --> 00:25:26.390

Joel Smith: One question, back to the duplicated items. Someone asked, can a list of the duplicated items that, have been removed be generated or displayed? So, can you get a view of, kind of, what duplicates you have before you remove them, or after?

179

00:25:26.390 --> 00:25:30.020

Allan Yarish: No, that's a good question. I think what you'd want to do

180

00:25:30.320 --> 00:25:37.479

Allan Yarish: The way I would handle that is, especially with a small table like this, is just copy this table.

181

00:25:37.770 --> 00:25:40.149

Allan Yarish: Maybe put it over here somewhere.

182

00:25:40.690 --> 00:25:48.719

Allan Yarish: Like that, and... would... What, make it a little bit more visible? Like that.

183

00:25:49.270 --> 00:25:55.179

Allan Yarish: So now, if you... Do, remove duplicates over here.

184

00:25:59.300 --> 00:26:07.460

Allan Yarish: Unfortunately, you'd have to go through and sort of manual figure out which ones got taken out, but it looks like clearly this... here's one that got taken out.

185

00:26:07.690 --> 00:26:14.930

Allan Yarish: Here's one, no, that didn't get taken out,

186

00:26:17.990 --> 00:26:20.510

Allan Yarish: Yeah, here's another one that got taken out.

187

00:26:20.850 --> 00:26:23.830

Allan Yarish: There were 3 of them, and it was probably that one, yeah.

188

00:26:24.020 --> 00:26:26.040

Joel Smith: Nice, I've just seen,

189

00:26:26.360 --> 00:26:34.780

Joel Smith: a comment in the chat as well. Someone mentioned you can use conditional formatting to see duplicated values before you remove anything, so I think that.

190

00:26:34.780 --> 00:26:35.380

Allan Yarish: Yes.

191

00:26:35.870 --> 00:26:37.239

Allan Yarish: Yeah, that.

192

00:26:37.240 --> 00:26:39.440

Joel Smith: tab. You could try that too.

193

00:26:39.440 --> 00:26:43.959

Allan Yarish: That... that is possible, and... Let's go back to...

194

00:26:45.230 --> 00:26:48.609

Allan Yarish: Conditional formatting would... but it requires...

195

00:26:48.900 --> 00:26:51.019

Allan Yarish: Some thought and effort to,

196

00:26:51.250 --> 00:27:01.180

Allan Yarish: You'd have to... you do it... Conditional formatting is... You create a rule.

197

00:27:01.940 --> 00:27:07.070

Allan Yarish: covering that data, you'd use a formula, and you'd have to write a formula, is,

198

00:27:07.070 --> 00:27:07.610

Joel Smith: duplicated.

199

00:27:07.610 --> 00:27:15.099

Allan Yarish: Cell A4 equal to A5? Is cell B4 equal to B5, etc, etc, etc, and then it would highlight them,

200

00:27:15.430 --> 00:27:19.600

Allan Yarish: I'll just do it, let's say,

201

00:27:39.250 --> 00:27:43.460

Allan Yarish: This would... this would just highlight...

202

00:27:54.160 --> 00:27:56.320

Allan Yarish: So, this highlights.

203

00:27:56.610 --> 00:27:57.070

Joel Smith: Oh, yeah.

204

00:27:57.070 --> 00:28:08.319

Allan Yarish: any ones where there appear to be duplicates. I did something weird with the formula, because it's highlighting ones that aren't... I don't know why it's highlighting that last row, but...

205

00:28:08.680 --> 00:28:13.920

Allan Yarish: You can do it, but it takes a little time and effort to figure out exactly the formula that you want to,

206

00:28:14.150 --> 00:28:15.659

Allan Yarish: That you want to apply.

207

00:28:16.490 --> 00:28:18.110

Joel Smith: Okay, perfect.

208

00:28:18.250 --> 00:28:29.940

Joel Smith: Great. Well, I'm afraid that is all we have time for. This is a speedy old event. So, I... yeah, I just want to, first of all,

209

00:28:29.940 --> 00:28:52.650

Joel Smith: give Alan a huge virtual round of applause, some really useful tips that I'm sure we'll all be taking away into our day-to-day work. And, most importantly, thank you to you, the audience at home, for dialing in and being so invested in your learning. I can see the emoji reactions flying up, all of the virtual applause, and everyone's thanks in the comments, which is lovely to see.

210

00:28:52.650 --> 00:29:00.810

Joel Smith: The team's gonna pull up a quick poll, just to find out how you felt about today's session. This will help us to continue improving our content, and we appreciate your responses.

211

00:29:00.900 --> 00:29:24.360

Joel Smith: Before you go, and while you are responding to that, I just want to share a few more events that you can register for to join us again soon, and if you're not available to be there live, remember you can always register for the event in order to get sent the recording to catch up whenever suits you. Those have just gone into the chat. Next Wednesday, we have Hacks for Working with Gemini, with guidance directly from a Google public sector expert.

212

00:29:24.480 --> 00:29:44.180

Joel Smith: On 13th of August, we have top lessons for Successful Government AI projects on what public servants can learn from real government AI pilots' successes and failures. And at

the end of August, we have our next AI Skills Lab edition, AI for Data Visualization, where you can learn how to visualize and clean the data sets that you have cleaned up after today's event.

213

00:29:44.260 --> 00:29:57.990

Joel Smith: Okay, that is really all we have time for today. We hope you found it helpful, and we want to say, again, a huge thank you to Alan, our speaker, and to everyone who made today happen, and to you for taking time out of your busy days.

214

00:29:57.990 --> 00:30:07.460

Joel Smith: Keep an eye out for the email coming your way with today's recording and resources, and don't forget to comment under my post. We hope to see you at another Apolitical event very soon.

215

00:30:07.460 --> 00:30:10.159

Joel Smith: Thanks very much, everybody. Have a great day. Goodbye.