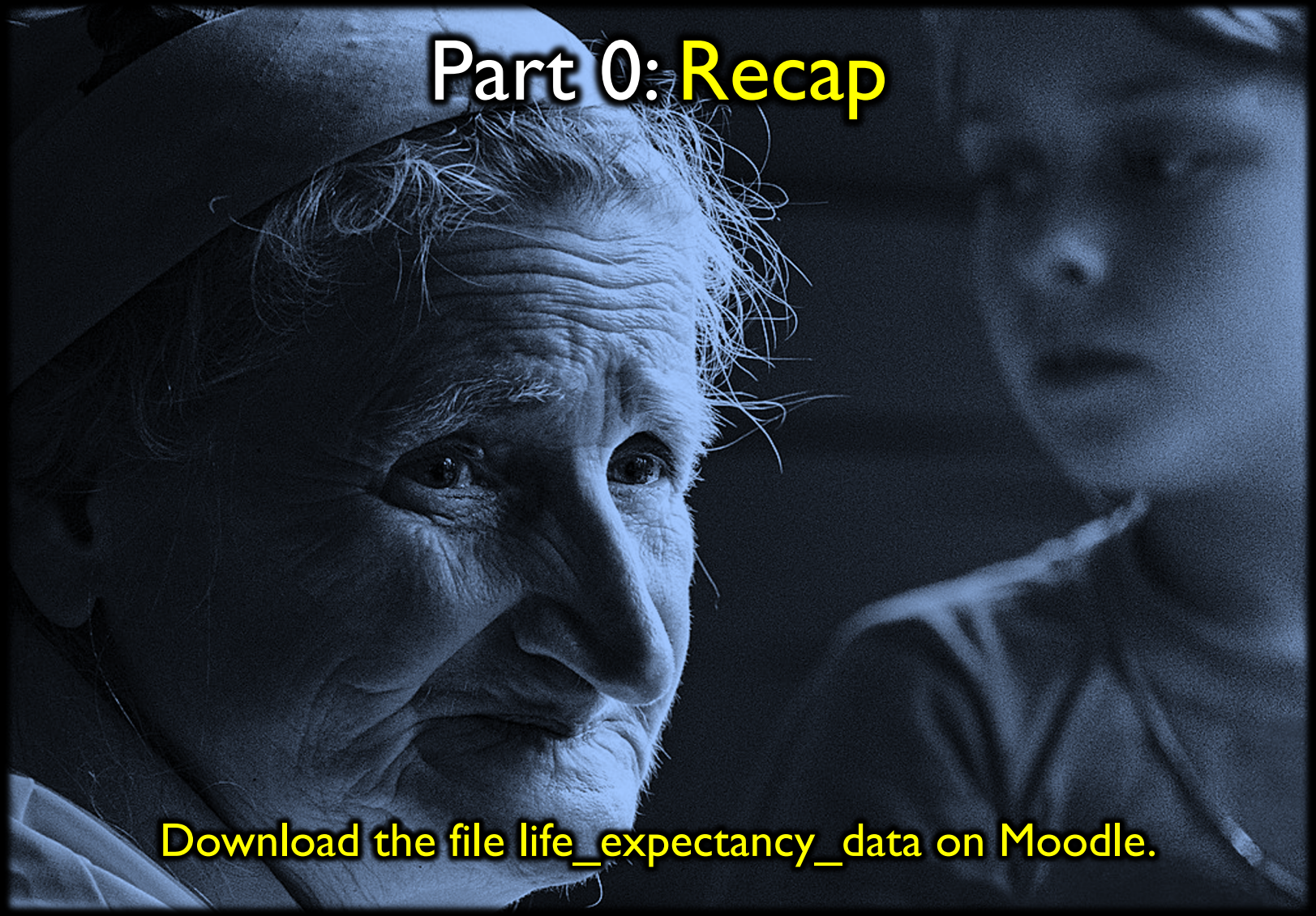
Part 0: Recap



Download the file life_expectancy_data on Moodle.

Based on what you have seen in the module so far, how would you explore and communicate this data with visuals and statistics?

IMAGE: Vitalii Shmorgun

Life Expectancy Data

Ward	Local_Authority	male_life_expectancy	female_life_expectancy
Belsize	Camden	84.2	88.2
Bloomsbury	Camden	81.8	86.2
Camden Town with Primrose Hill	Camden	82.7	86.5
Cantelowes	Camden	78.2	85.3
Fortune Green	Camden	82.7	87.8
Frognal and Fitzjohns	Camden	84.8	88.0
Gospel Oak	Camden	80.2	86.1
Hampstead Town	Camden	84.0	90.7
Haverstock	Camden	78.1	85.5
Highgate	Camden	83.3	86.6
Holborn and Covent Garden	Camden	79.4	90.2
Kentish Town	Camden	76.7	80.4
Kilburn	Camden	76.8	80.3
Kings Cross	Camden	80.1	88.0
Regents Park	Camden	77.1	84.7
St Pancras and Somers Town	Camden	77.4	81.7
Swiss Cottage	Camden	81.4	87.5
West Hampstead	Camden	80.7	85.9
Bowes	Enfield	83.0	85.1
Bush Hill Park	Enfield	81.7	84.1
Chase	Enfield	77.3	81.4
Cockfosters	Enfield	82.9	86.1

1. Use excel to complete a table of the following summary statistics:

	Male	Female
	all_data	all_data
quantity		
mean		
median		
min		
max		
range		
LQ		
UQ		
IQR		
variance		
st. dev.		

Summary Statistics

	Male	Female
quantity	57	57
mean	81.2	85.5
median	81.0	85.5
min	74.8	78.6
max	91.9	92.9
range	17.1	14.3
LQ	78.4	84.1
UQ	83.3	87.2
IQR	4.9	3.1
variance	11.1	7.3
st. dev.	3.3	2.7

2. Use the formula for Tukey fences to calculate the boundaries for outliers

	Male	Female
quantity	57	57
mean	81.2	85.5
median	81.0	85.5
min	74.8	78.6
max	91.9	92.9
range	17.1	14.3
LQ	78.4	84.1
UQ	83.3	87.2
IQR	4.9	3.1
variance	11.1	7.3
st. dev.	3.3	2.7
lo-tukey	71.0	79.4
hi-tukey	90.6	91.8

3. Use the COUNTIF formula on excel to identify the number of outliers

	Male	Female
quantity	57	57
mean	81.2	85.5
median	81.0	85.5
min	74.8	78.6
max	91.9	92.9
range	17.1	14.3
LQ	78.4	84.1
UQ	83.3	87.2
IQR	4.9	3.1
variance	11.1	7.3
st. dev.	3.3	2.7
lo-tukey	71.0	79.4
hi-tukey	90.6	91.8
lo-outliers	0	1
hi-outliers	1	1

4. Repeat your summary statistics for each region

	Male				Female			
	all_data	Camden	Enfield	K&C	all_data	Camden	Enfield	K&C
quantity	57	18	21	18	57	18	21	18
mean	81.2	80.5	80.1	83.0	85.5	86.1	84.1	86.6
median	81.0	80.4	80.9	83.5	85.5	86.3	84.4	86.4
min	74.8	76.7	76.1	74.8	78.6	80.3	78.6	81.9
max	91.9	84.8	83.5	91.9	92.9	90.7	87.2	92.9
range	17.1	8.1	7.4	17.1	14.3	10.4	8.6	11.0
LQ	78.4	78.2	78.8	79.9	84.1	85.3	82.5	85.3
UQ	83.3	82.7	81.7	86.3	87.2	87.9	85.4	88.0
IQR	4.9	4.6	2.9	6.4	3.1	2.6	2.9	2.7
lo-tukey	71.0	71.3	74.4	70.2	79.4	81.4	78.2	81.3
hi-tukey	90.6	89.6	86.2	96.0	91.8	91.9	89.7	92.0
lo-outliers	0	0	0	0	1	2	0	0
hi-outliers	1	0	0	0	1	0	0	1
variance	11.1	7.0	4.8	17.6	7.3	8.0	4.0	6.4
st. dev.	3.3	2.6	2.2	4.2	2.7	2.8	2.0	2.5

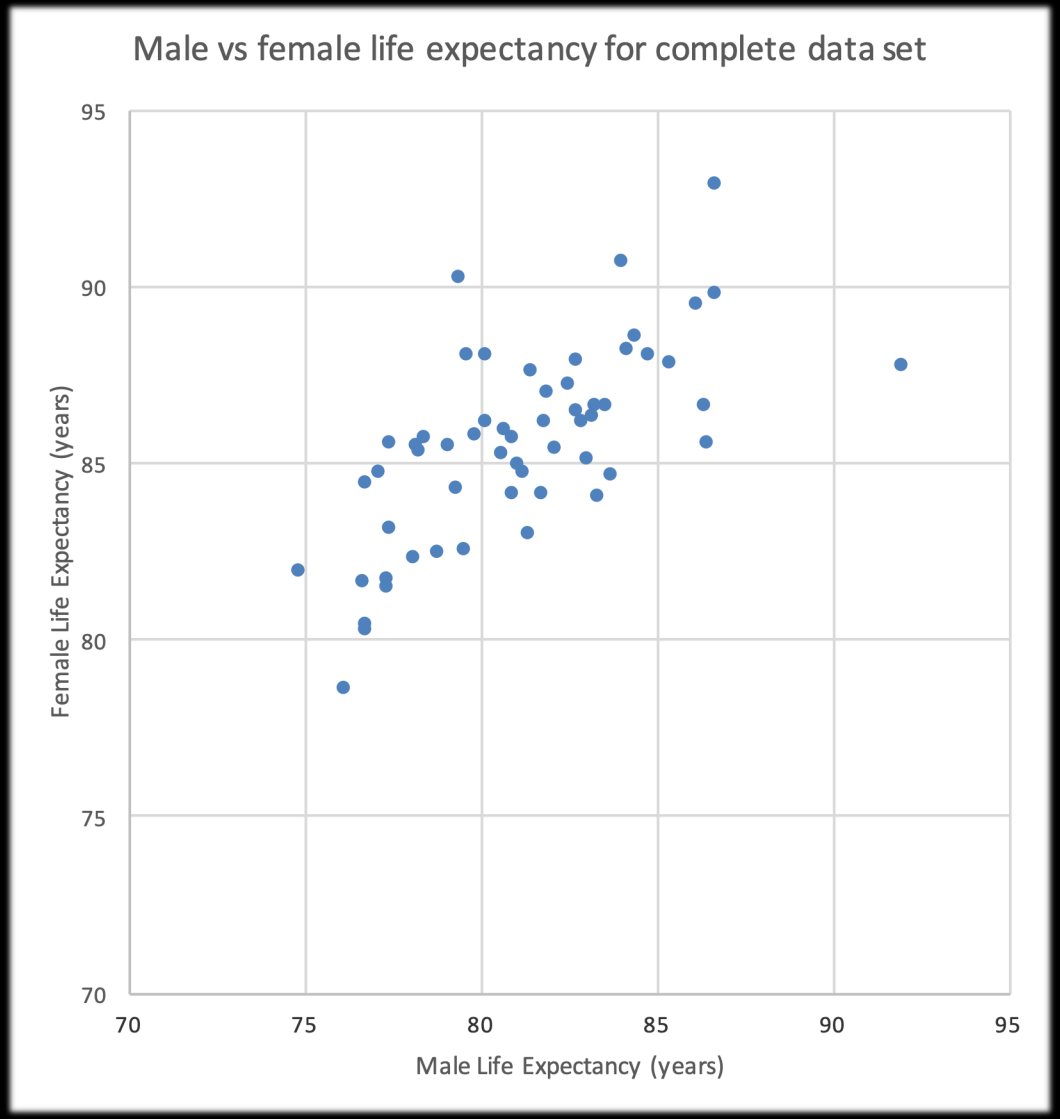
5. What do you notice?
What do you wonder?

	Male				Female			
	all_data	Camden	Enfield	K&C	all_data	Camden	Enfield	K&C
quantity	57	18	21	18	57	18	21	18
mean	81.2	80.5	80.1	83.0	85.5	86.1	84.1	86.6
median	81.0	80.4	80.9	83.5	85.5	86.3	84.4	86.4
min	74.8	76.7	76.1	74.8	78.6	80.3	78.6	81.9
max	91.9	84.8	83.5	91.9	92.9	90.7	87.2	92.9
range	17.1	8.1	7.4	17.1	14.3	10.4	8.6	11.0
LQ	78.4	78.2	78.8	79.9	84.1	85.3	82.5	85.3
UQ	83.3	82.7	81.7	86.3	87.2	87.9	85.4	88.0
IQR	4.9	4.6	2.9	6.4	3.1	2.6	2.9	2.7
lo-tukey	71.0	71.3	74.4	70.2	79.4	81.4	78.2	81.3
hi-tukey	90.6	89.6	86.2	96.0	91.8	91.9	89.7	92.0
lo-outliers	0	0	0	0	1	2	0	0
hi-outliers	1	0	0	0	1	0	0	1
variance	11.1	7.0	4.8	17.6	7.3	8.0	4.0	6.4
st. dev.	3.3	2.6	2.2	4.2	2.7	2.8	2.0	2.5

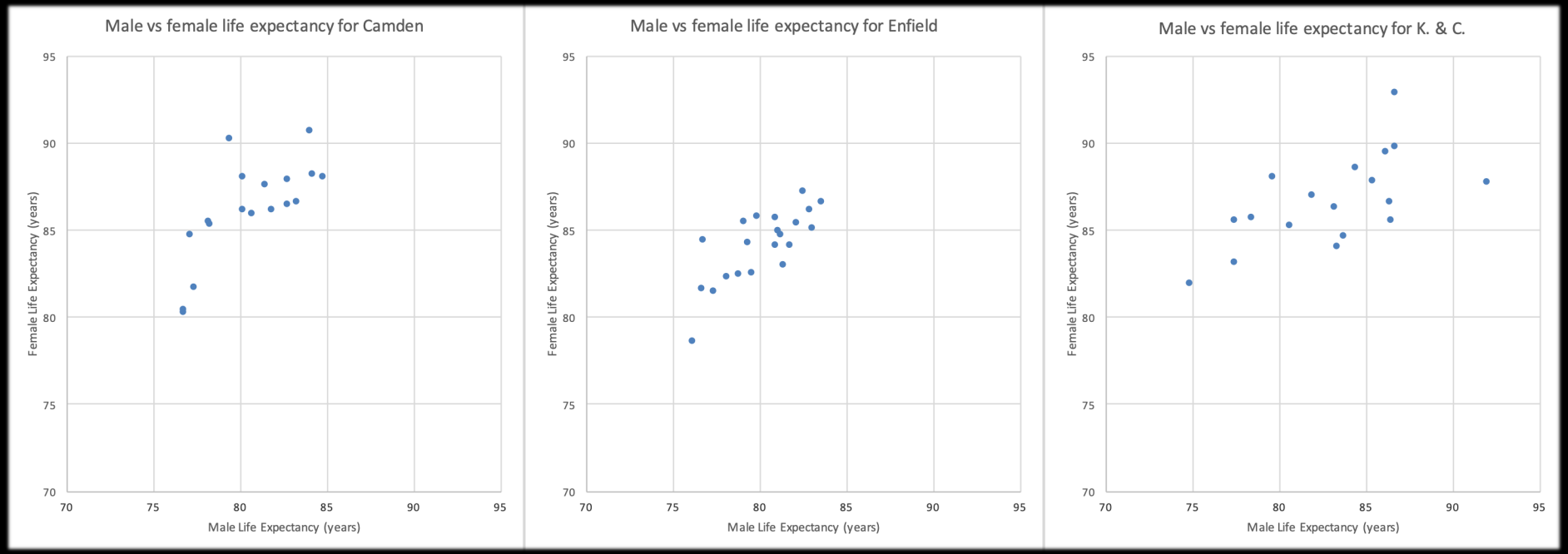
Here's one thing worth noticing:

	Male				Female			
	all_data	Camden	Enfield	K&C	all_data	Camden	Enfield	K&C
quantity	57	18	21	18	57	18	21	18
mean	81.2	80.5	80.1	83.0	85.5	86.1	84.1	86.6
median	81.0	80.4	80.9	83.5	85.5	86.3	84.4	86.4
min	74.8	76.7	76.1	74.8	78.6	80.3	78.6	81.9
max	91.9	84.8	83.5	91.9	92.9	90.7	87.2	92.9
range	17.1	8.1	7.4	17.1	14.3	10.4	8.6	11.0
LQ	78.4	78.2	78.8	79.9	84.1	85.3	82.5	85.3
UQ	83.3	82.7	81.7	86.3	87.2	87.9	85.4	88.0
IQR	4.9	4.6	2.9	6.4	3.1	2.6	2.9	2.7
lo-tukey	71.0	71.3	74.4	70.2	79.4	81.4	78.2	81.3
hi-tukey	90.6	89.6	86.2	96.0	91.8	91.9	89.7	92.0
lo-outliers	0	0	0	0	1	2	0	0
hi-outliers	1	0	0	0	1	0	0	1
variance	11.1	7.0	4.8	17.6	7.3	8.0	4.0	6.4
st. dev.	3.3	2.6	2.2	4.2	2.7	2.8	2.0	2.5

6. Make another sheet in excel and create the following Scatter Plot

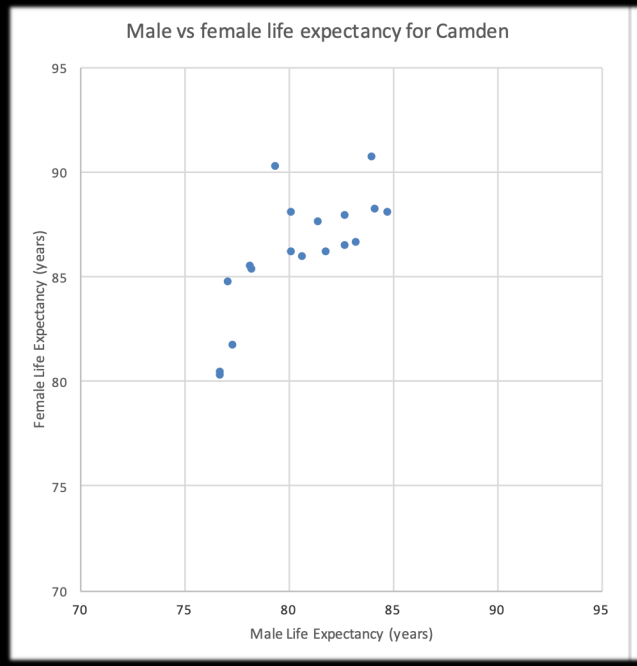


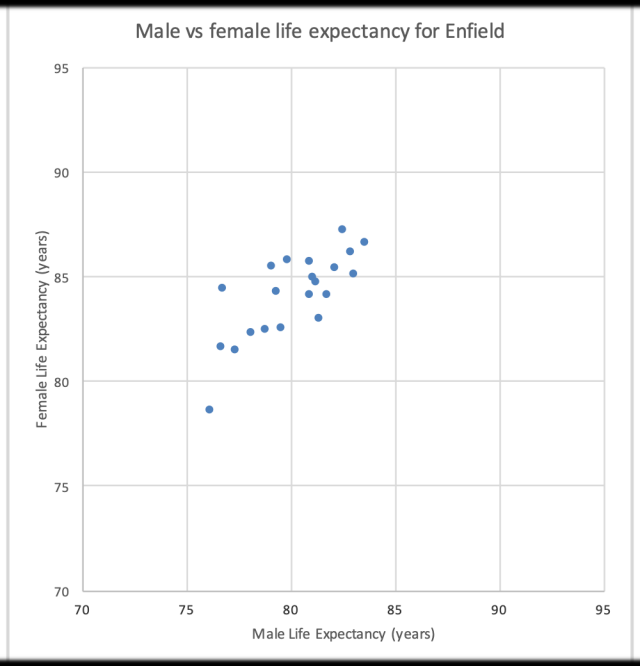
7. Create the following scatter plots

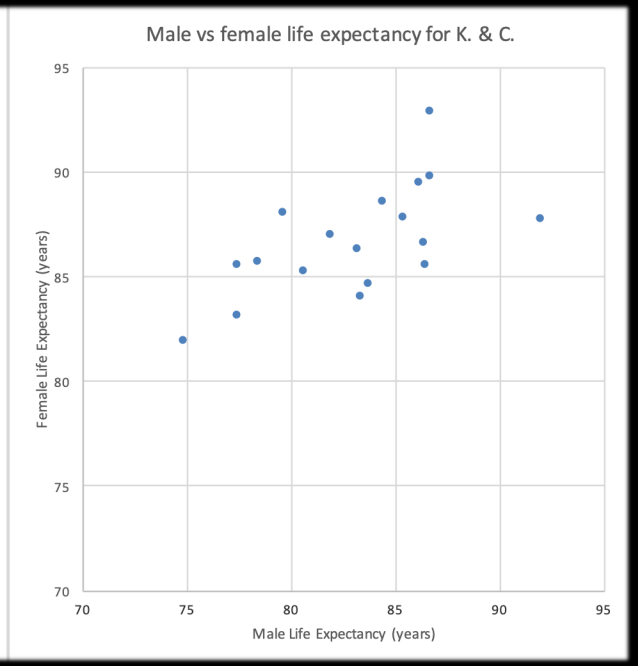


Which approach is better to see the data? The separate plots? Or all together?

7. Create the following scatter plots

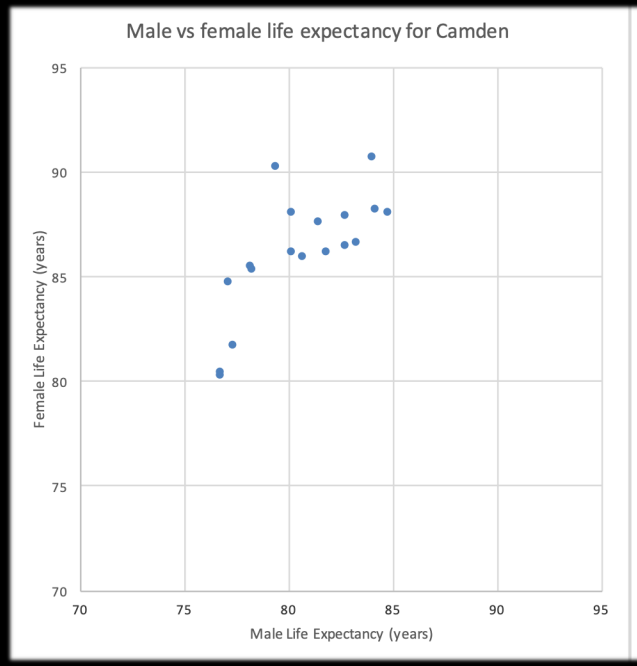


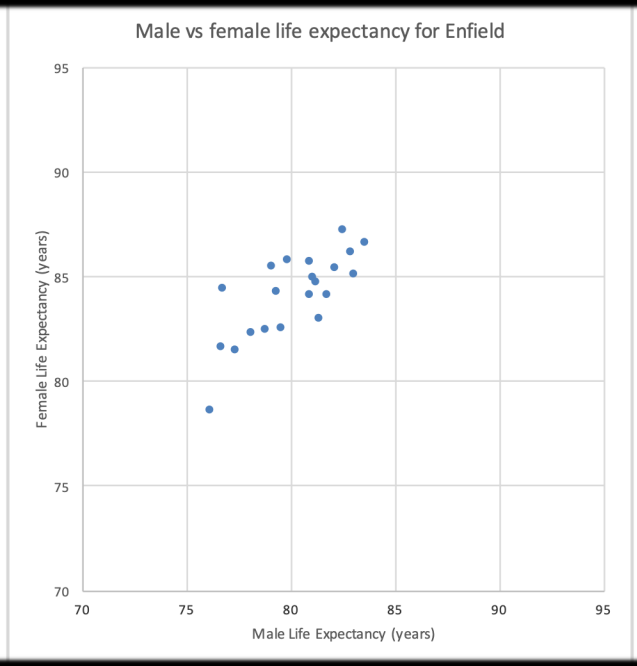


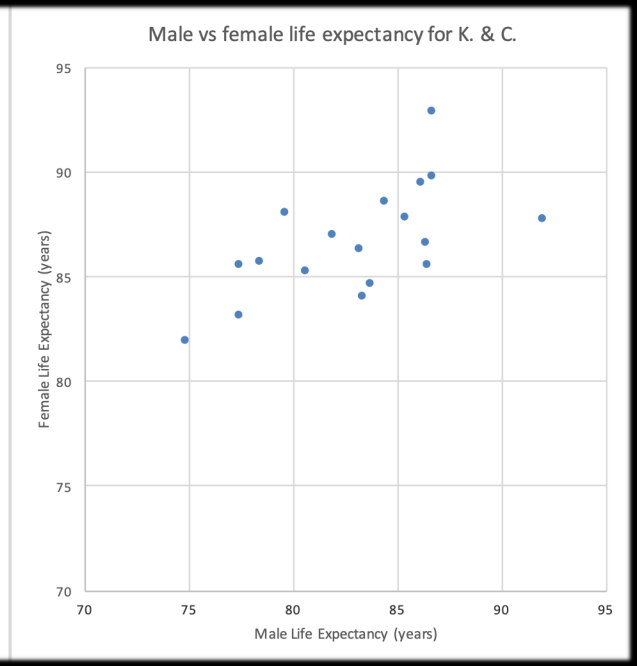


How else might we present these plots?

Now time to think about outliers



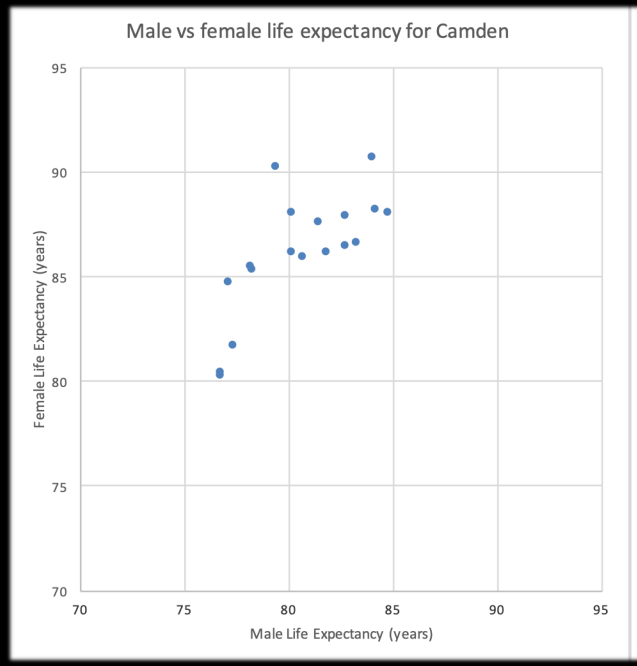


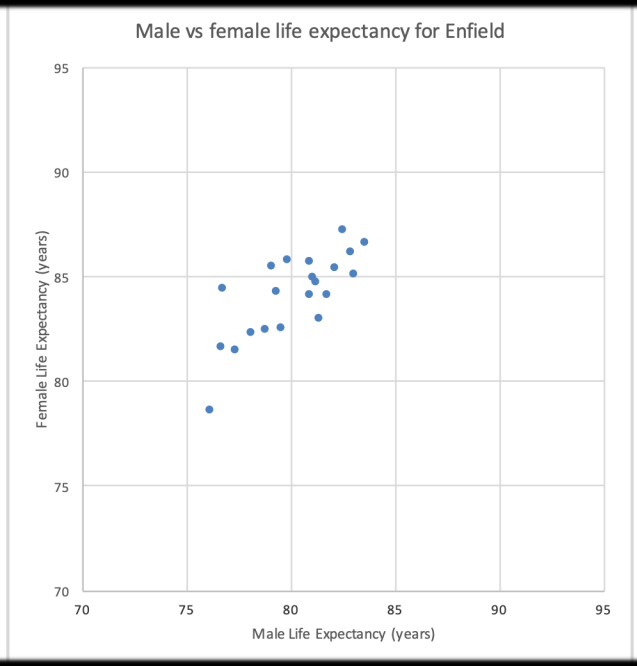


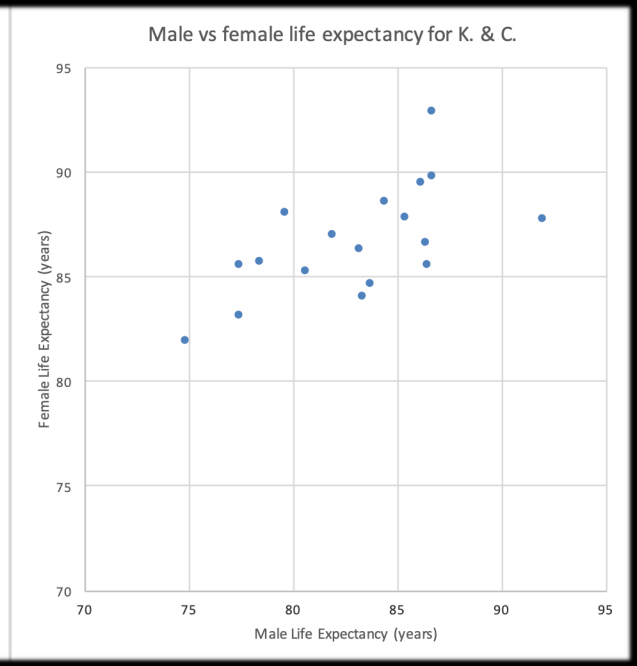
There are two in Camden and one in K&C

	Male				Female			
	all_data	Camden	Enfield	K&C	all_data	Camden	Enfield	K&C
quantity	57	18	21	18	57	18	21	18
mean	81.2	80.5	80.1	83.0	85.5	86.1	84.1	86.6
median	81.0	80.4	80.9	83.5	85.5	86.3	84.4	86.4
min	74.8	76.7	76.1	74.8	78.6	80.3	78.6	81.9
max	91.9	84.8	83.5	91.9	92.9	90.7	87.2	92.9
range	17.1	8.1	7.4	17.1	14.3	10.4	8.6	11.0
LQ	78.4	78.2	78.8	79.9	84.1	85.3	82.5	85.3
UQ	83.3	82.7	81.7	86.3	87.2	87.9	85.4	88.0
IQR	4.9	4.6	2.9	6.4	3.1	2.6	2.9	2.7
lo-tukey	71.0	71.3	74.4	70.2	79.4	81.4	78.2	81.3
hi-tukey	90.6	89.6	86.2	96.0	91.8	91.9	89.7	92.0
lo-outliers	0	0	0	0	1	2	0	0
hi-outliers	1	0	0	0	1	0	0	1
variance	11.1	7.0	4.8	17.6	7.3	8.0	4.0	6.4
st. dev.	3.3	2.6	2.2	4.2	2.7	2.8	2.0	2.5

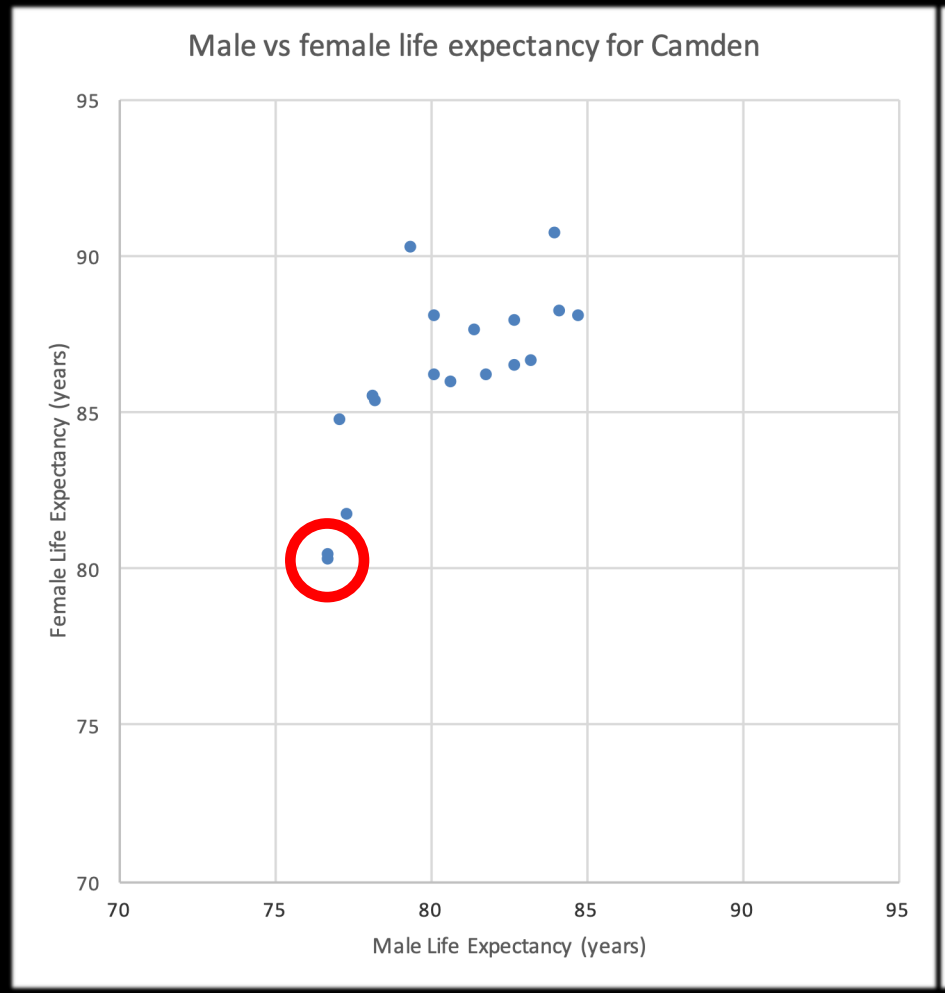
8. Which ones do you think they are?

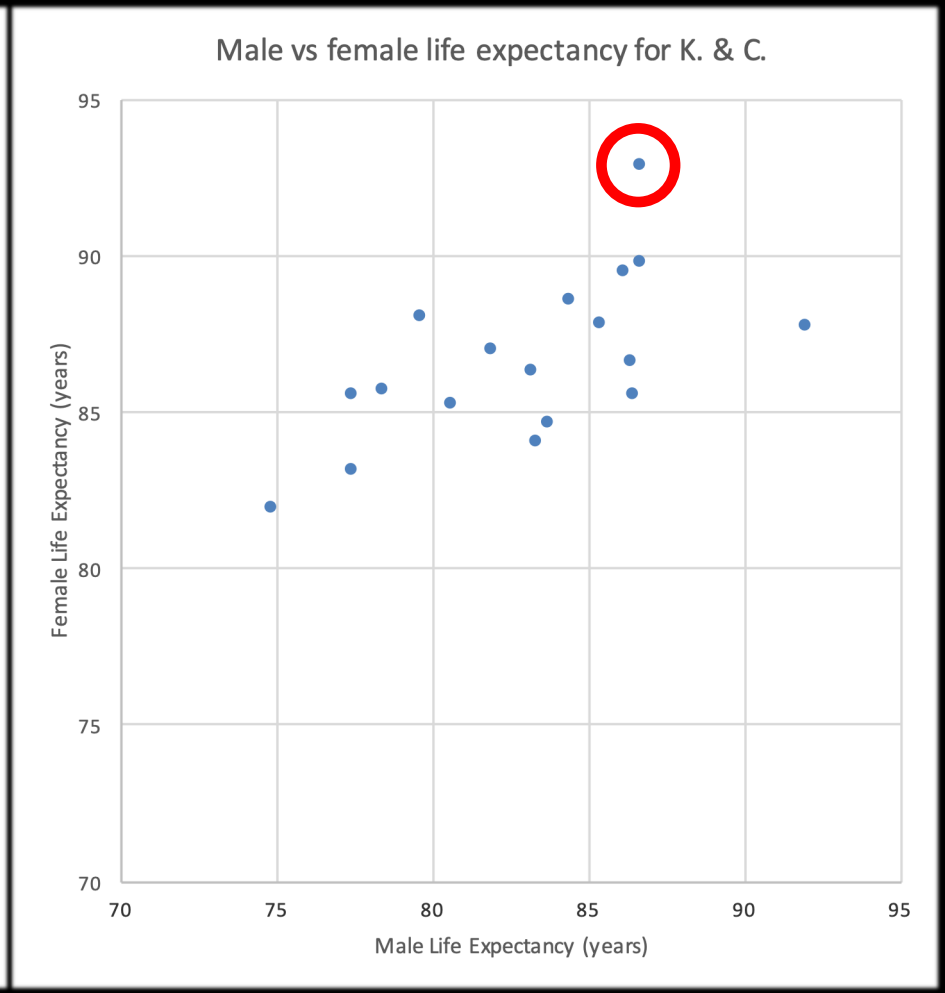






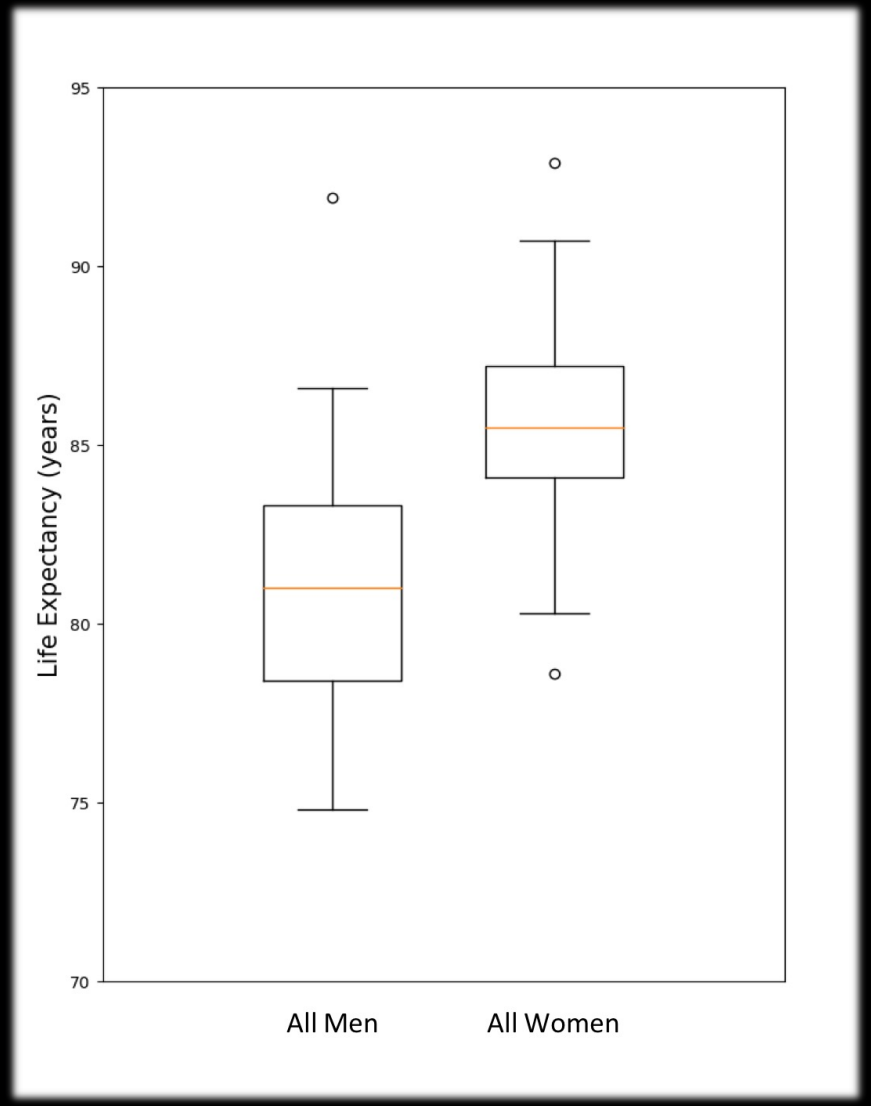
Outliers



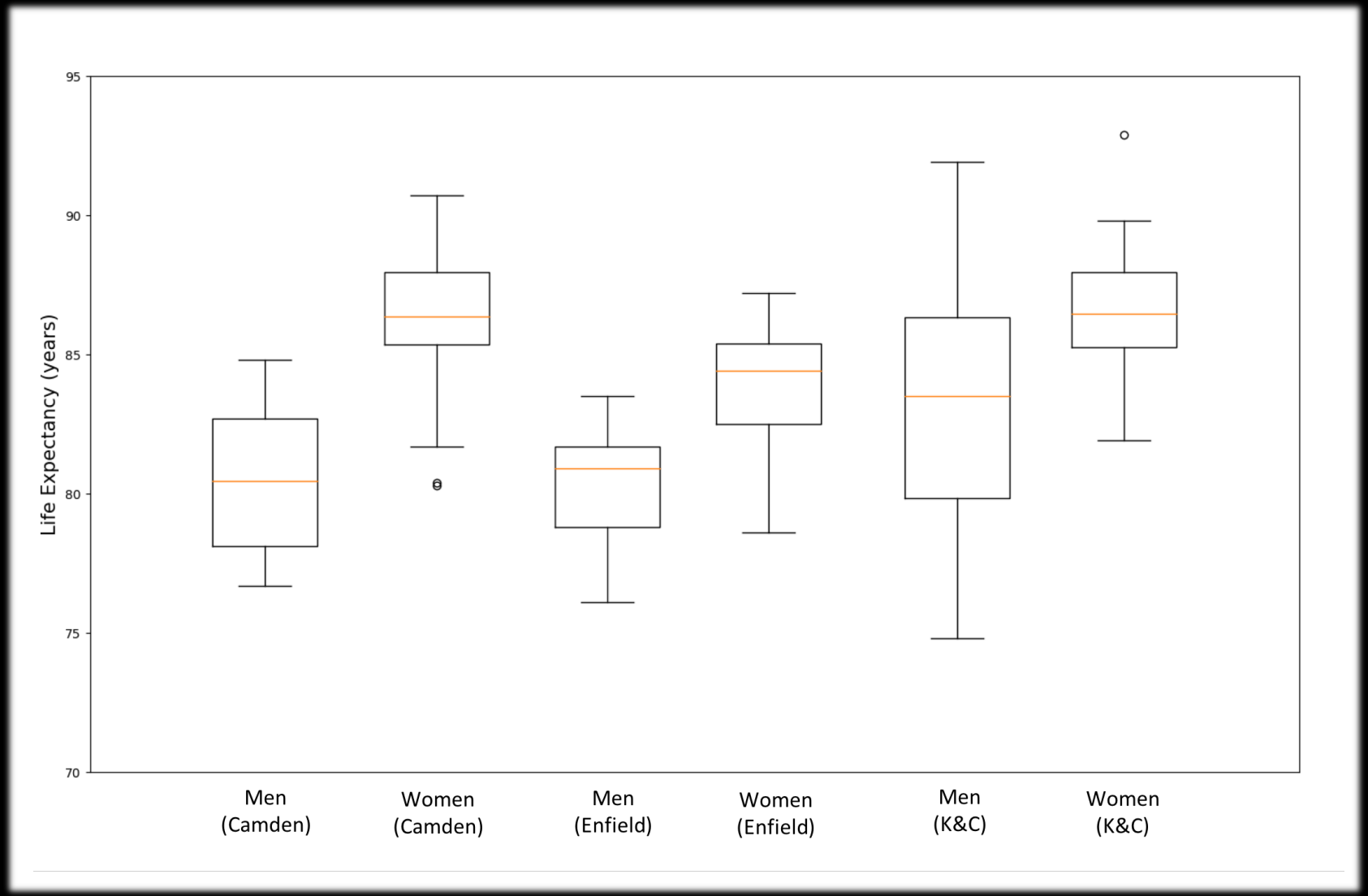


What do you think should be done about these outliers?

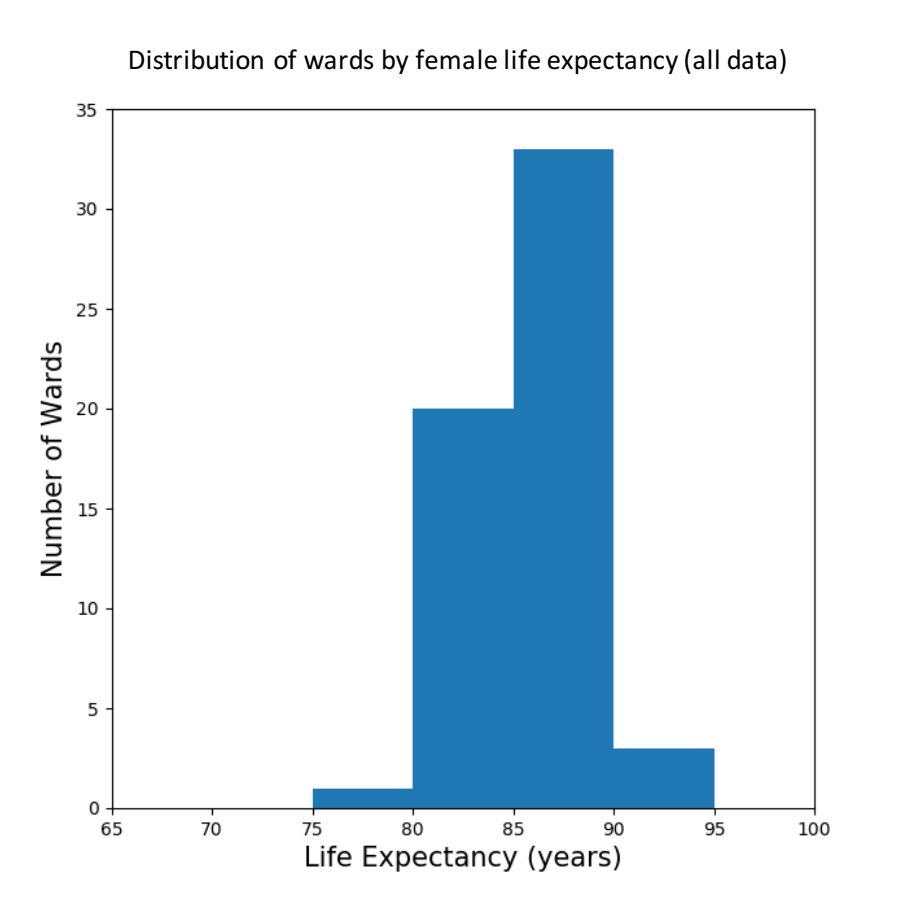
9. Create the following box plots on a new excel sheet



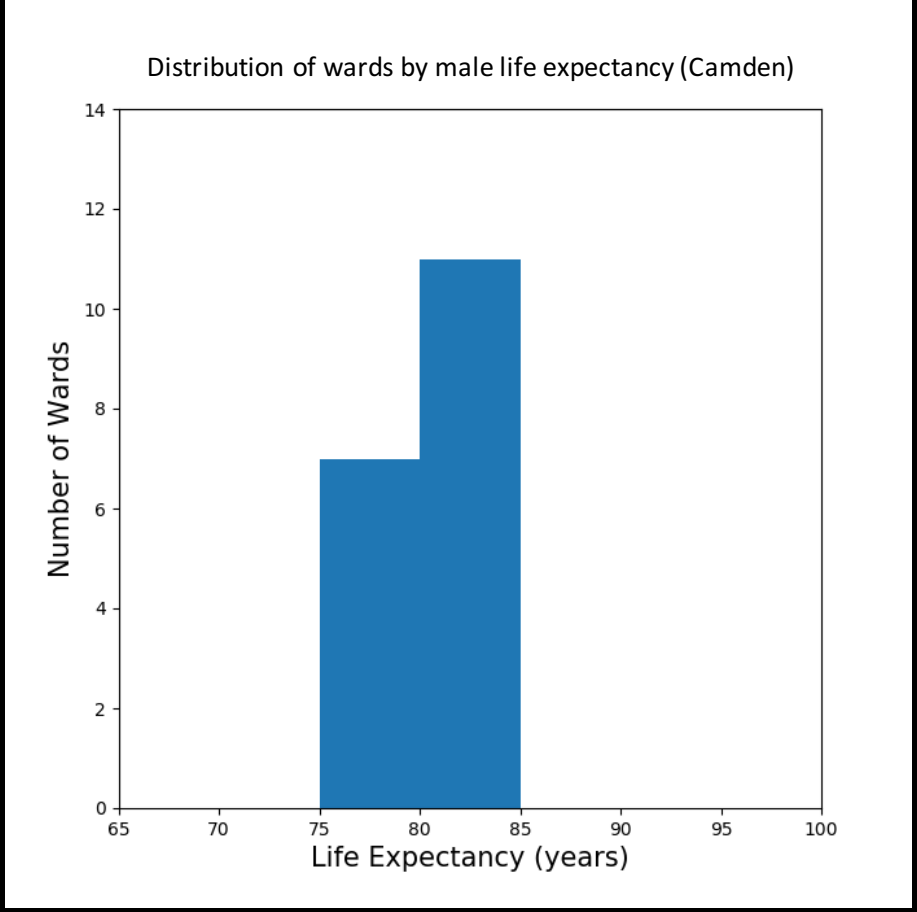
9. Create the following box plots on a new excel sheet

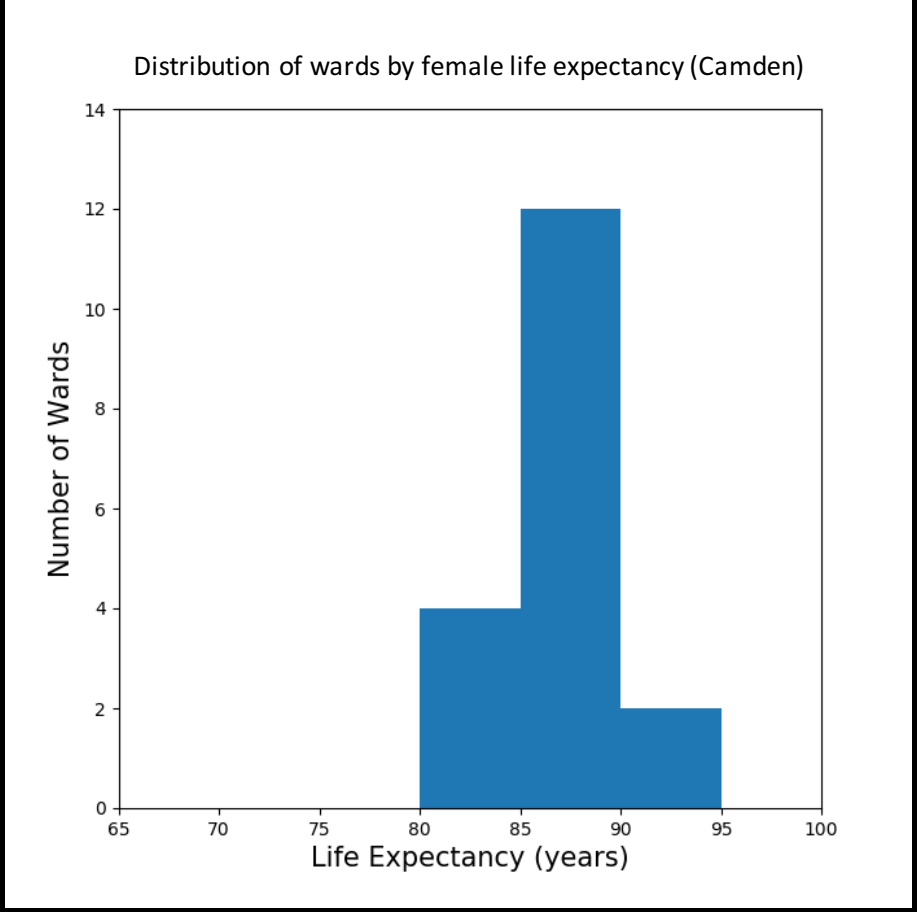


10. Now try to make these histograms on another new sheet

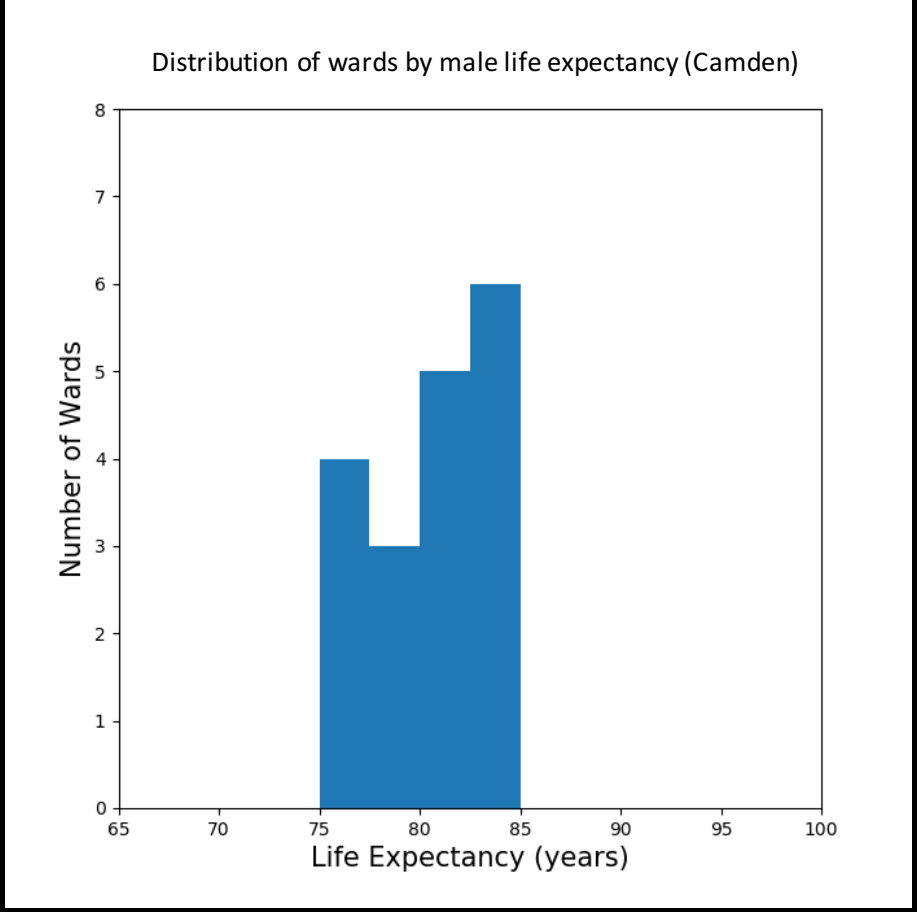


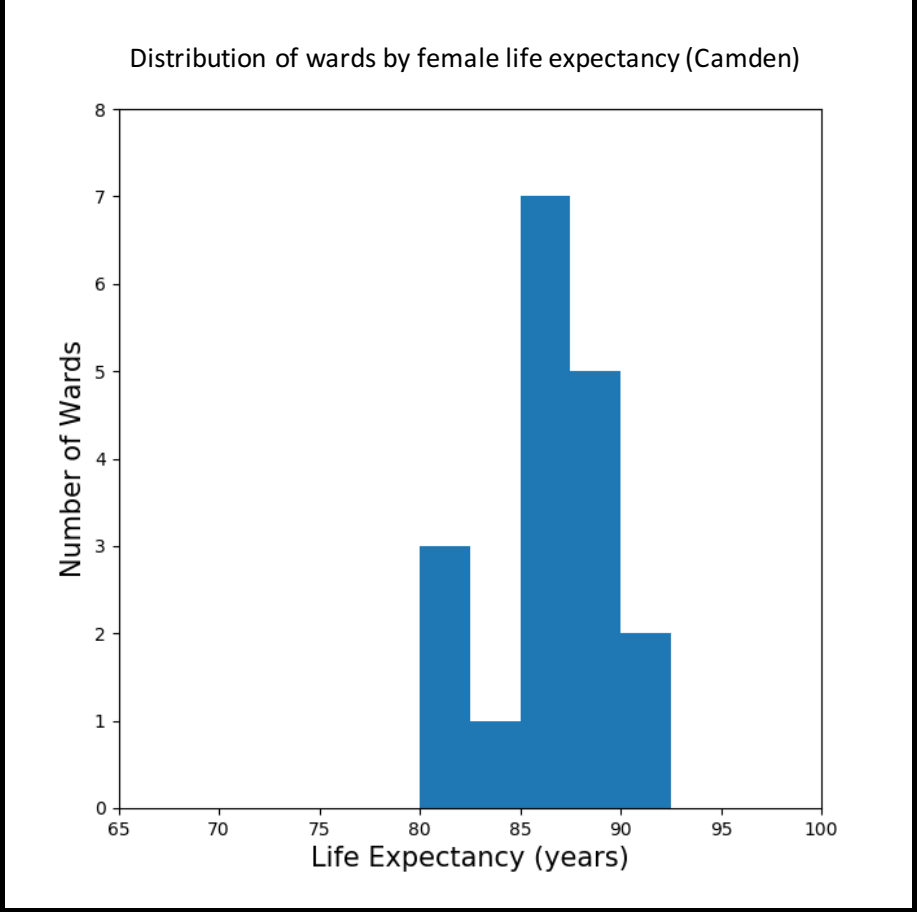
10. Now try to make these histograms on another new sheet





10. Now try to make these histograms on another new sheet





11. Now try to recreate the following correlation table

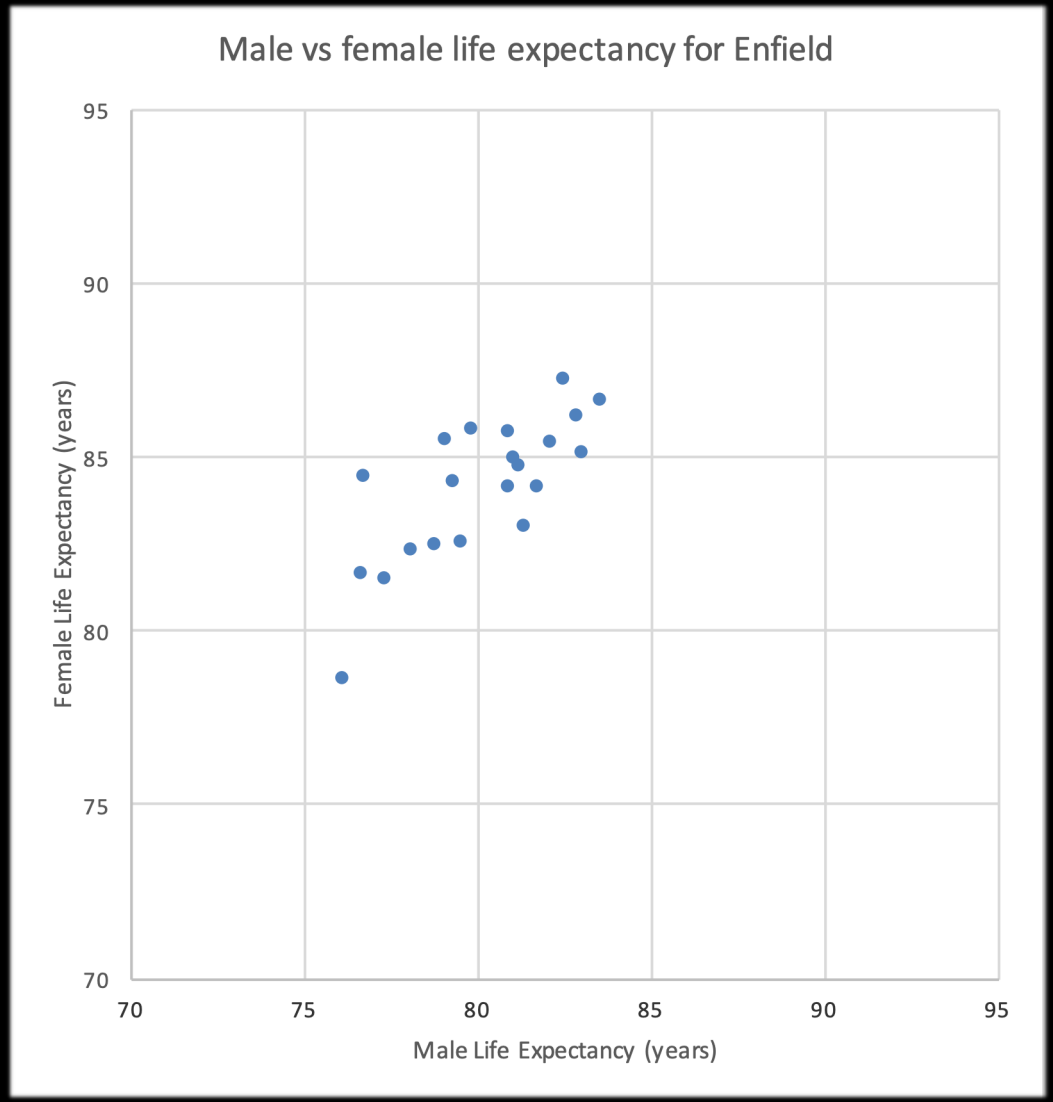
Correlations (M vs. F)	Pearson	Spearman
Overall	0.68	0.70
Camden	0.74	0.77
Enfield	0.76	0.72
Kensington and Chelsea	0.62	0.71

11. Now try to recreate the following correlation table

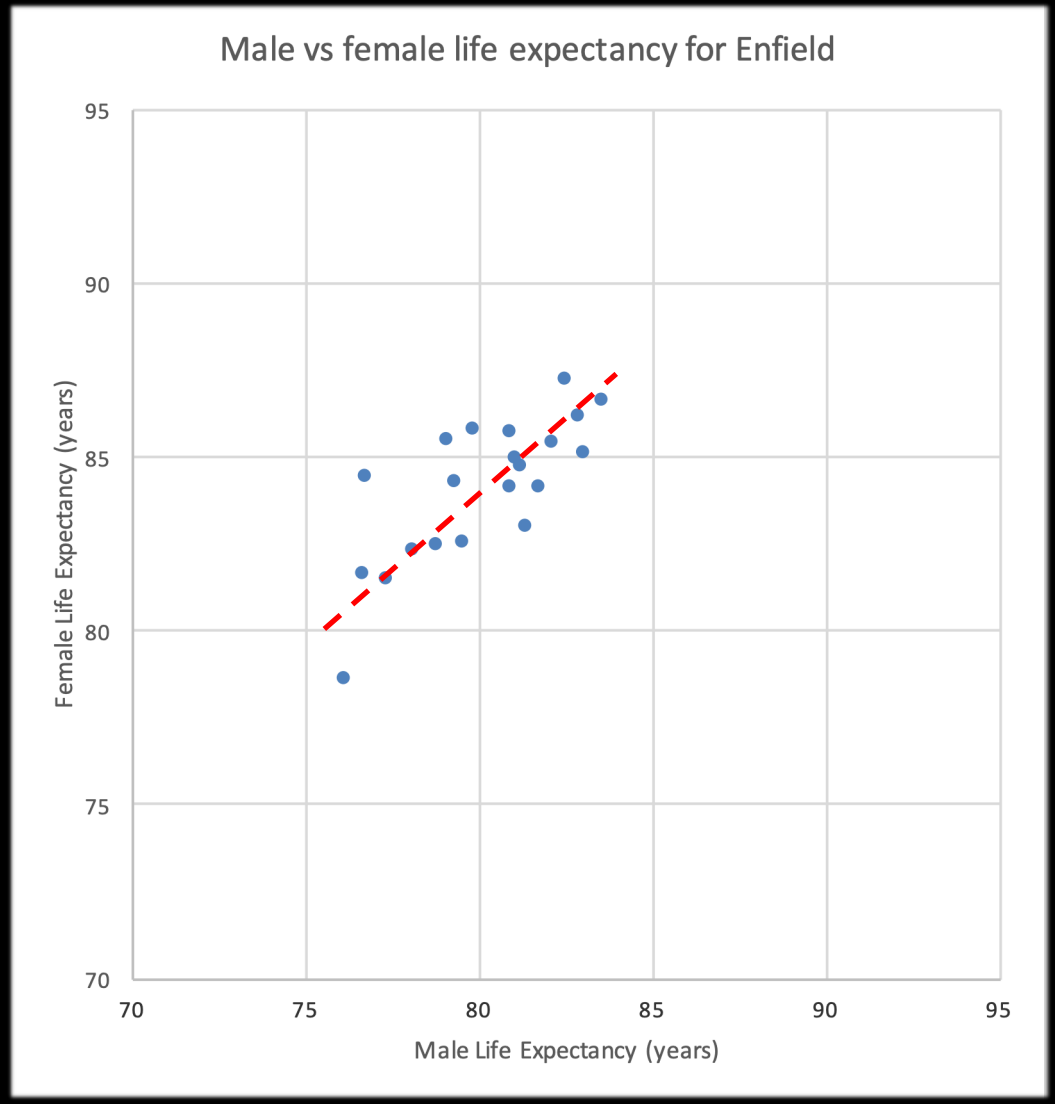
Correlations (M vs. F)	Pearson	Spearman
Overall	0.68	0.70
Camden	0.74	0.77
Enfield	0.76	0.72
Kensington and Chelsea	0.62	0.71

Which is more suitable for this data?

Linear Regression

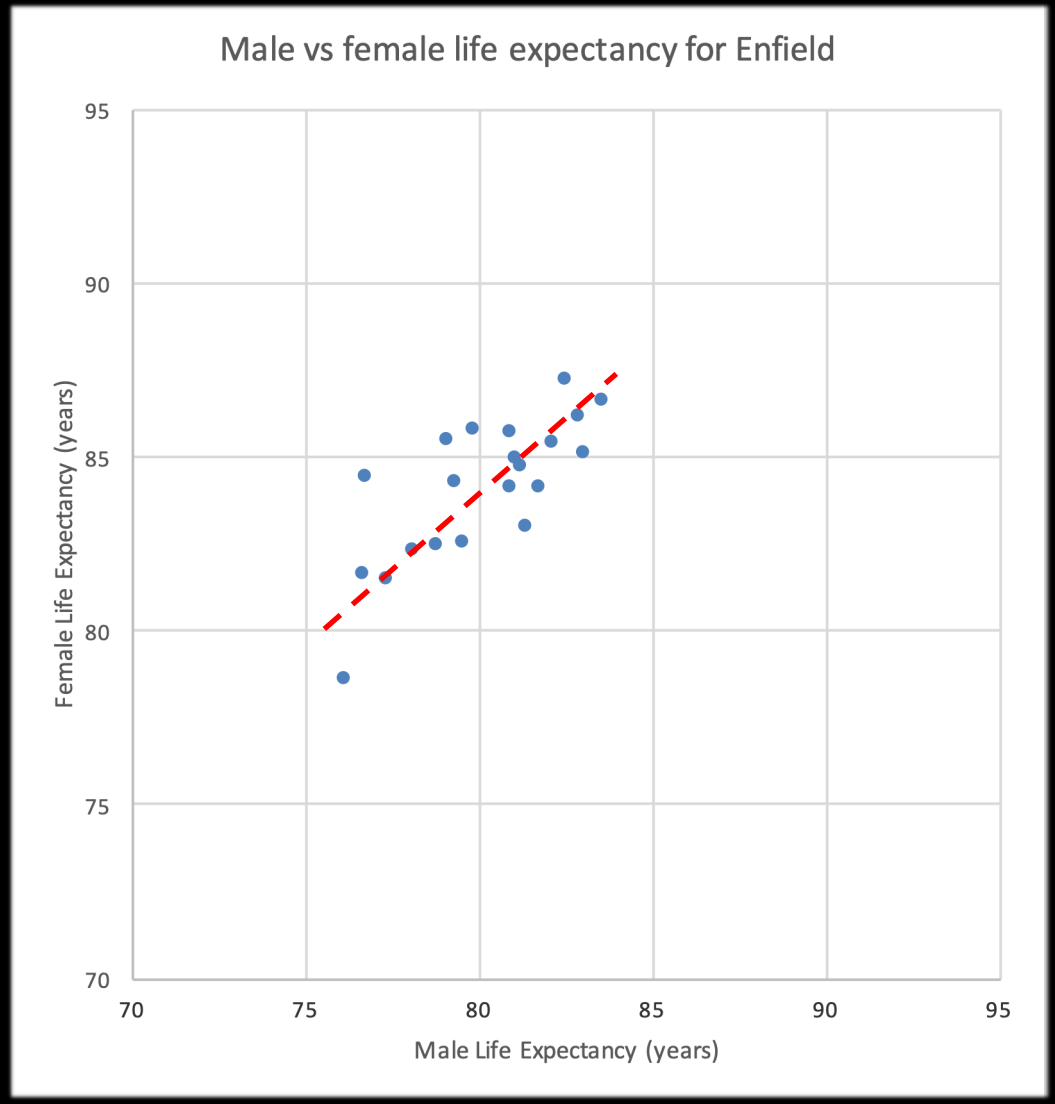


12. Linear Regression



What values do you expect the regression to give you?

12. Linear Regression



What values will the regression give you?

simple_regression_python.py

Simple Linear Regression

What to Report and Visualise

Fitted Equation:	$\hat{y} = mx + c$
p-value:	Is the result significant?
R-Squared Value:	Strength of relationship
Data Scatter Plot:	With fitted line
Residuals vs Fits Plot:	Consider LINE conditions

LINEAR REGRESSION

Necessary Conditions

Linear relationship exists

Independent errors

Normally distributed errors

Equal variance for all x values