2. Compute the standard *descriptive statistics* for the variable pm10.

a. Compare the Analysis ToolPack results to those of a few built-in functions.

Answer) They seem to be identical.

Column1			
		4	**=COUNTBLANK(A2:A851)**
Mean	93.40216315	93.4021631	**=AVERAGE(A2:A851)**
Standard Error	1.109286217	1.10928622	**=STDEV(A2:A851)/SQRT(COUNT(A2:A851))**
Median	91.1805	91.1805	**=MEDIAN(A2:A851)**
Mode	97	97	**=MODE(A2:A851)**
Standard Deviation	32.26478671	32.2647867	**=STDEV(A2:A851)**
Sample Variance	1041.016462	1041.01646	**=VAR(A2:A851)**
Kurtosis	47.32347917	47.3234792	**=KURT(A2:A851)**
Skewness	3.796381164	3.79638116	**=SKEW(A2:A851)**
Range	521	521	**=MAX(A2:A851)-MIN(A2:A851)**
Minimum	29	29	**=MIN(A2:A851)**
Maximum	550	550	**=MAX(A2:A851)**
Sum	79018.23002	79018.23	**=SUM(A2:A851)**
Count	846	846	**=COUNT(A2:A851)**
	-1.0548E-165		

b. What do the statistics tell you about the data?

Answer) There are 4 blank values. The mean is 93.40. The median is 91.18. Minimum is 29. Maximum is 550. There are 846 values.

3. Create a variable measuring GDP per capita in U.S. dollars.

a. Use the exchange rate added to the data.

b. What is the mean, median, s.d., min and maximum value of this variable?

Answer)

Column1	
Mean	2527725.151
Standard Error	137686.5465
Median	1104874.75
Mode	#N/A
Standard Deviation	4014218.145
Sample Variance	1.61139E+13
Kurtosis	15.89288044
Skewness	3.65691803
Range	31156872.82
Minimum	102084.8319
Maximum	31258957.65
Sum	2148566378
Count	850
	-1.0548E-165

4. Replicate the histogram below. You're welcome to ignore titles but not labels.

a. Replicate the histogram for the variable manuf_share.

b. Create a second histogram of PM10 excluding the outlier.

c. Create a histogram of the natural log of PM10 (again excluding the outlier).

5. Describe the shape of the distribution of 2007 rainfall across Chinese cities.

a. Use another histogram. What is the correct number of observations?

Answer) 850 (as expected from the data description).

b. What then is the suggested number of bin according to Module.A.2?

Answer)

Formula 1:	Number of bins = √n		29.1547595
Answer:		**29**	
Formula 2:	Number of bins = 10ln(n)/ln(10)		29.2941893
Answer:		**29**	

6. Use pol_chn.xlsx to review PivotTables from Module A.1. Answer the following.

a. How many different years and cities are there in the data?

Answer) Years: 10, Cities: 85

Row Labels	Count of city_id
2	10
3	10
5	10
7	10
17	10
18	10
19	10
20	10
29	10
32	10

Row Labels	Count of year
2003	85
2004	85
2005	85
2006	85
2007	85
2008	85
2009	85
2010	85
2011	85
2012	85

b. Compute the average pollution level and manufacturing share for each year.

Row Labels	Avg. Pollution	Avg. Manufacturing
2003	154.429436	50.70905873
2004	119.53213	52.12400018
2005	108.151623	49.47199988
2006	99.3884704	50.26082375
2007	94.471894	50.64105876
2008	88.0103882	51.23047056
2009	81.4590116	49.55399991
2010	73.3728823	50.06188234
2011	64.8263061	50.86141158
2012	48.2549136	49.96

c. Make the following pivot table of the crosstabulation of the two variables.

Row Labels	Average of pm10	Average of manuf_share
2003	154.429436	50.70905873
2004	119.53213	52.12400018
2005	108.151623	49.47199988
2006	99.3884704	50.26082375
2007	94.471894	50.64105876
2008	88.0103882	51.23047056
2009	81.4590116	49.55399991
2010	73.3728823	50.06188234
2011	64.8263061	50.86141158
2012	48.2549136	49.96
Grand Total	**93.4021631**	**50.48747057**

7. Agree or disagree. Support your argument with additional evidence from the data.

a. A decline in pollution is unlikely to occur evenly across China. Why?

Answer) Disagree. Looking at the table below, I observe a declining trend in pollution in every city. Thus, a decline in pollution does seem to occur evenly across different cities in China.

Average of pm10	Column Labels					
Row Labels	**2**	**3**	**5**	**7**	**17**	**18**
2003	550	239	198	192	186	181
2004	128	128	128	127.00001	127.00001	127.00001
2005	113	113	112	112	112	112
2006	102	102	102	102	102	102
2007	97	97	97	97	97	97
2008	91	91	91	90.391998	90.246994	90
2009	85	85	85	85	85	85
2010	77.887001	77.844002	77.815002	77.202995	77	77
2011	68.758003	68.706993	68.681999	68.333	68.209999	68.121002
2012	59	59	59	58.525002	58	58
Grand Total	**137.1645004**	**106.0550995**	**101.8497001**	**100.9453005**	**100.2457003**	**99.7121012**

8. Recall Zheng and Kahn (2017) and the data in pol_chn.xlsx.

b. Add 1-2paragraphs summarizing the research question, data, & conclusion.

The research aims to answer how air pollution caused by urbanization in China affects quality of life and health in her cities, and its costs and consequences. The paper uses data which contain the following variables and their units:
Year, Unique ID for the Chinese cities, Air pollution level measured as PM10, Annual rainfall in 2007 in millimeters, City's longitude in degrees East, City's latitude in degrees North, 2007 temperature discomfort index, City's GDP in constant (2012) 10,000s of RMB, City population (nonagricultural) in 10,000s of people, Percent share of employment in manufacturing, Average years of schooling in 2000, Dummy variables for specific cities, and China/US Exchange Rate (Chinese Yuan to $1 USD).

There are 850 observations and 17 variables in the data. There are 2 identifiers for years and cities, 10 quantitative variables and 5 categorical variables. The paper concludes that pollution in China do cause health damages such as shorter life expectancy and loss of consumption, and thus impart economic burdens to the country.