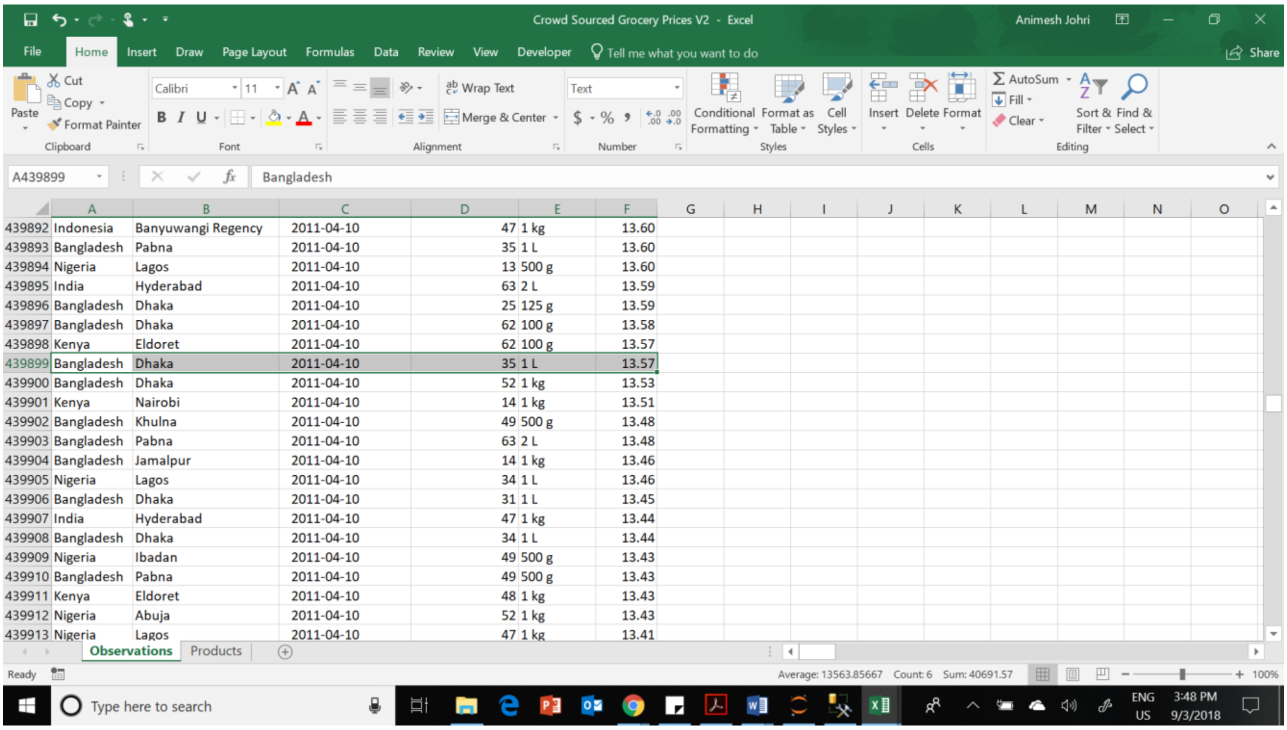
Exercise- Data Cleansing Exercise

Student Name: Animesh Johri **Student Id:**2021292753

Date: 09/07/2018

a) The value in cell D439899 is 35

Screenshot:



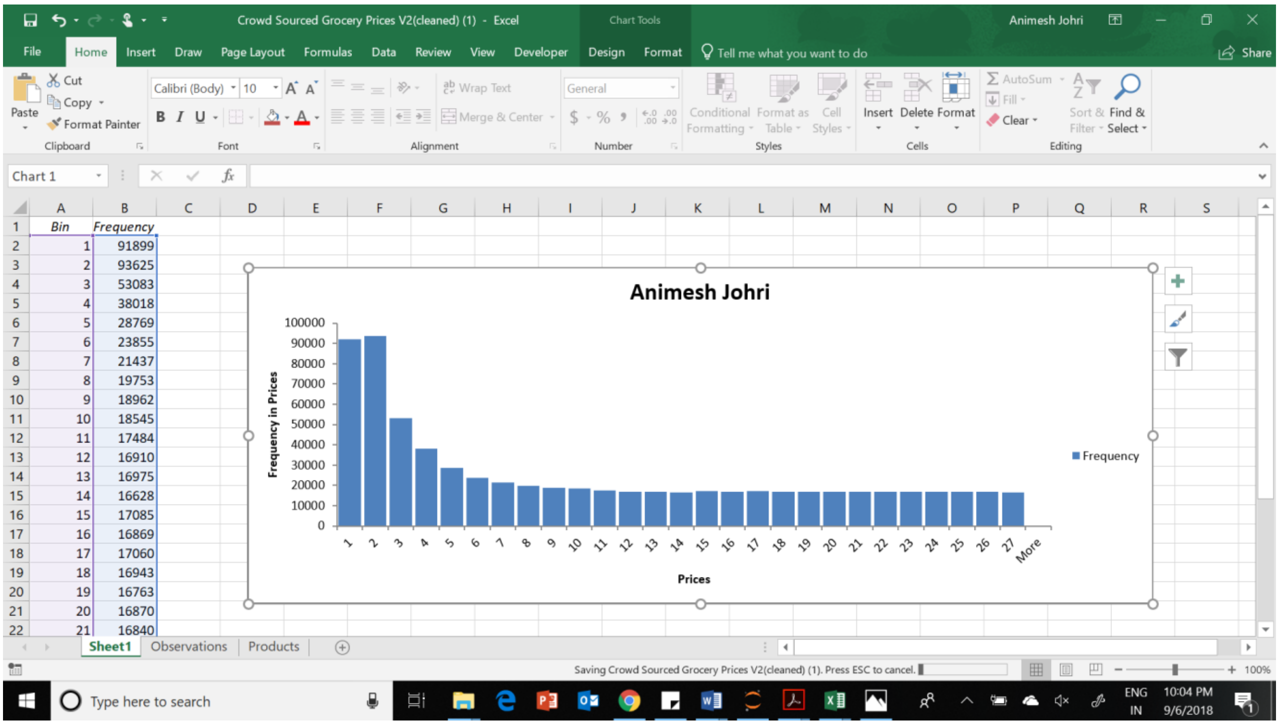
A439899 | Bangladesh

	A	B	C	D	E	F
439892	Indonesia	Banyuwangi Regency	2011-04-10	47	1 kg	13.60
439893	Bangladesh	Pabna	2011-04-10	35	1 L	13.60
439894	Nigeria	Lagos	2011-04-10	13	500 g	13.60
439895	India	Hyderabad	2011-04-10	63	2 L	13.59
439896	Bangladesh	Dhaka	2011-04-10	25	125 g	13.59
439897	Bangladesh	Dhaka	2011-04-10	62	100 g	13.58
439898	Kenya	Eldoret	2011-04-10	62	100 g	13.57
439899	Bangladesh	Dhaka	2011-04-10	35	1 L	13.57
439900	Bangladesh	Dhaka	2011-04-10	52	1 kg	13.53
439901	Kenya	Nairobi	2011-04-10	14	1 kg	13.51
439902	Bangladesh	Khulna	2011-04-10	49	500 g	13.48
439903	Bangladesh	Pabna	2011-04-10	63	2 L	13.48
439904	Bangladesh	Jamalpur	2011-04-10	14	1 kg	13.46
439905	Nigeria	Lagos	2011-04-10	34	1 L	13.46
439906	Bangladesh	Dhaka	2011-04-10	31	1 L	13.45
439907	India	Hyderabad	2011-04-10	47	1 kg	13.44
439908	Bangladesh	Dhaka	2011-04-10	34	1 L	13.44
439909	Nigeria	Ibadan	2011-04-10	49	500 g	13.43
439910	Bangladesh	Pabna	2011-04-10	49	500 g	13.43
439911	Kenya	Eldoret	2011-04-10	48	1 kg	13.43
439912	Nigeria	Abuja	2011-04-10	52	1 kg	13.43
439913	Nigeria	Lagos	2011-04-10	47	1 kg	13.41

Average: 13563.85667 Count: 6 Sum: 40691.57

b) **The methodology chosen for imputation:** I made a histogram based on the price column, the histogram was skewed, therefore I chose to impute the median of the price column in the blank cells. To impute -> I selected the whole price column using "ctrl + shift + down arrow" -> I clicked "ctrl + H" to find and replace -> I found all the blank cells within the price column and replaced it with the median value of 8.91 (To find median, I selected all the values in price column and used the formula "=median" over the price column)

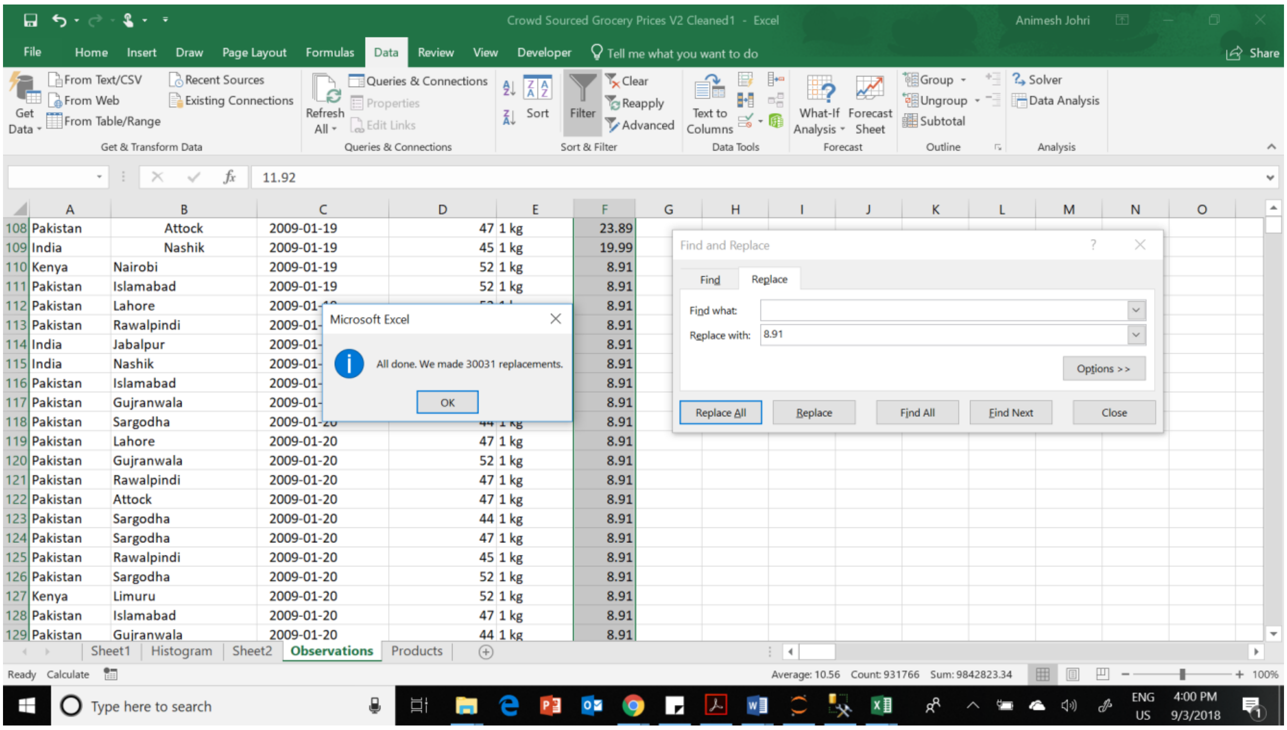
Screenshot of the histogram:



	A	B
1	*Bin*	*Frequency*
2	1	91899
3	2	93625
4	3	53083
5	4	38018
6	5	28769
7	6	23855
8	7	21437
9	8	19753
10	9	18962
11	10	18545
12	11	17484
13	12	16910
14	13	16975
15	14	16628
16	15	17085
17	16	16869
18	17	17060
19	18	16943
20	19	16763
21	20	16870
22	21	16840

Saving Crowd Sourced Grocery Prices V2(cleaned) (1). Press ESC to cancel.

c) A screenshot of the table with the imputed values:



11.92

	A	B	C	D	E	F
108	Pakistan	Attock	2009-01-19	47	1 kg	23.89
109	India	Nashik	2009-01-19	45	1 kg	19.99
110	Kenya	Nairobi	2009-01-19	52	1 kg	8.91
111	Pakistan	Islamabad	2009-01-19	52	1 kg	8.91
112	Pakistan	Lahore	2009-01-			8.91
113	Pakistan	Rawalpindi	2009-01-			8.91
114	India	Jabalpur	2009-01-			8.91
115	India	Nashik	2009-01-			8.91
116	Pakistan	Islamabad	2009-01-			8.91
117	Pakistan	Gujranwala	2009-01-			8.91
118	Pakistan	Sargodha	2009-01-20	44	1 kg	8.91
119	Pakistan	Lahore	2009-01-20	47	1 kg	8.91
120	Pakistan	Gujranwala	2009-01-20	52	1 kg	8.91
121	Pakistan	Rawalpindi	2009-01-20	47	1 kg	8.91
122	Pakistan	Attock	2009-01-20	47	1 kg	8.91
123	Pakistan	Sargodha	2009-01-20	44	1 kg	8.91
124	Pakistan	Sargodha	2009-01-20	47	1 kg	8.91
125	Pakistan	Rawalpindi	2009-01-20	45	1 kg	8.91
126	Pakistan	Sargodha	2009-01-20	52	1 kg	8.91
127	Kenya	Limuru	2009-01-20	52	1 kg	8.91
128	Pakistan	Islamabad	2009-01-20	47	1 kg	8.91
129	Pakistan	Gujranwala	2009-01-20	44	1 kg	8.91

Microsoft Excel

All done. We made 30031 replacements.

OK

Find and Replace

Find | Replace

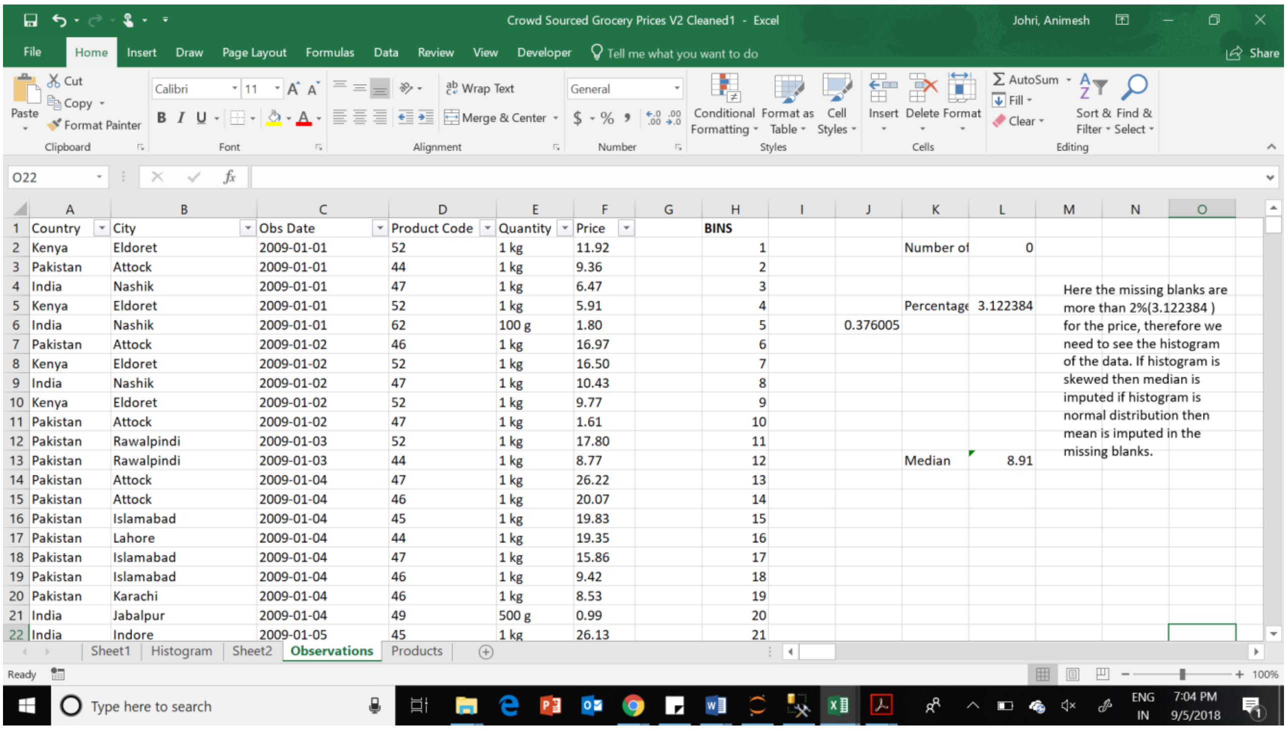
Find what:

Replace with: 8.91

Options >> | Replace All | Replace | Find All | Find Next | Close

Average: 10.56 Count: 931766 Sum: 9842823.34

d) A screenshot of the table after initial cleaning:



	A	B	C	D	E	F	G	H	I	J	K	L
1	Country	City	Obs Date	Product Code	Quantity	Price		BINS				
2	Kenya	Eldoret	2009-01-01	52	1 kg	11.92		1			Number of	0
3	Pakistan	Attock	2009-01-01	44	1 kg	9.36		2				
4	India	Nashik	2009-01-01	47	1 kg	6.47		3				
5	Kenya	Eldoret	2009-01-01	52	1 kg	5.91		4			Percentage	3.122384
6	India	Nashik	2009-01-01	62	100 g	1.80		5		0.376005		
7	Pakistan	Attock	2009-01-02	46	1 kg	16.97		6				
8	Kenya	Eldoret	2009-01-02	52	1 kg	16.50		7				
9	India	Nashik	2009-01-02	47	1 kg	10.43		8				
10	Kenya	Eldoret	2009-01-02	52	1 kg	9.77		9				
11	Pakistan	Attock	2009-01-02	47	1 kg	1.61		10				
12	Pakistan	Rawalpindi	2009-01-03	52	1 kg	17.80		11				
13	Pakistan	Rawalpindi	2009-01-03	44	1 kg	8.77		12			Median	8.91
14	Pakistan	Attock	2009-01-04	47	1 kg	26.22		13				
15	Pakistan	Attock	2009-01-04	46	1 kg	20.07		14				
16	Pakistan	Islamabad	2009-01-04	45	1 kg	19.83		15				
17	Pakistan	Lahore	2009-01-04	44	1 kg	19.35		16				
18	Pakistan	Islamabad	2009-01-04	47	1 kg	15.86		17				
19	Pakistan	Islamabad	2009-01-04	46	1 kg	9.42		18				
20	Pakistan	Karachi	2009-01-04	46	1 kg	8.53		19				
21	India	Jabalpur	2009-01-04	49	500 g	0.99		20				
22	India	Indore	2009-01-05	45	1 kg	26.13		21				

Here the missing blanks are more than 2%(3.122384) for the price, therefore we need to see the histogram of the data. If histogram is skewed then median is imputed if histogram is normal distribution then mean is imputed in the missing blanks.

e) The total number of rows (including header) present in the data sheet are 961733

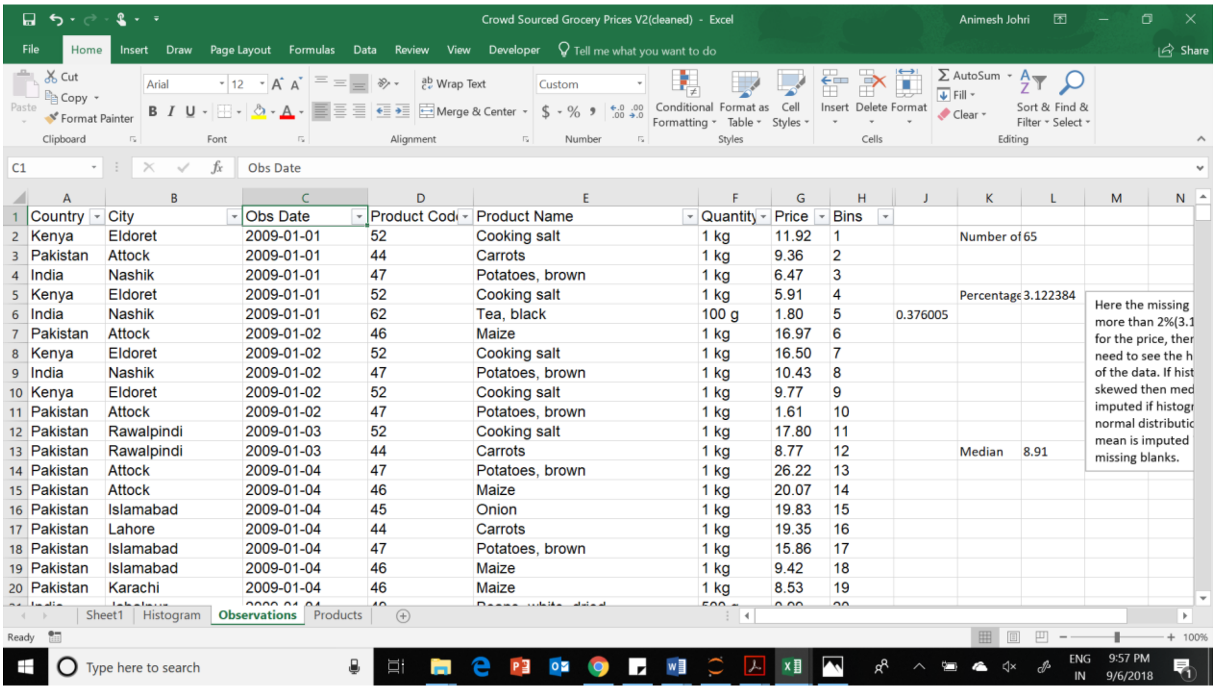
f) Countries with spell check error:

1. Brazil (wrong spelling: Brazel)
2. Indonesia (wrong spelling: Indonseia)
3. Philippines (wrong spelling: Phillippines).

The countries with duplicate entries are:

1. India
2. Bangladesh
3. Kenya
4. Pakistan

g. Screenshot showing the changes made:



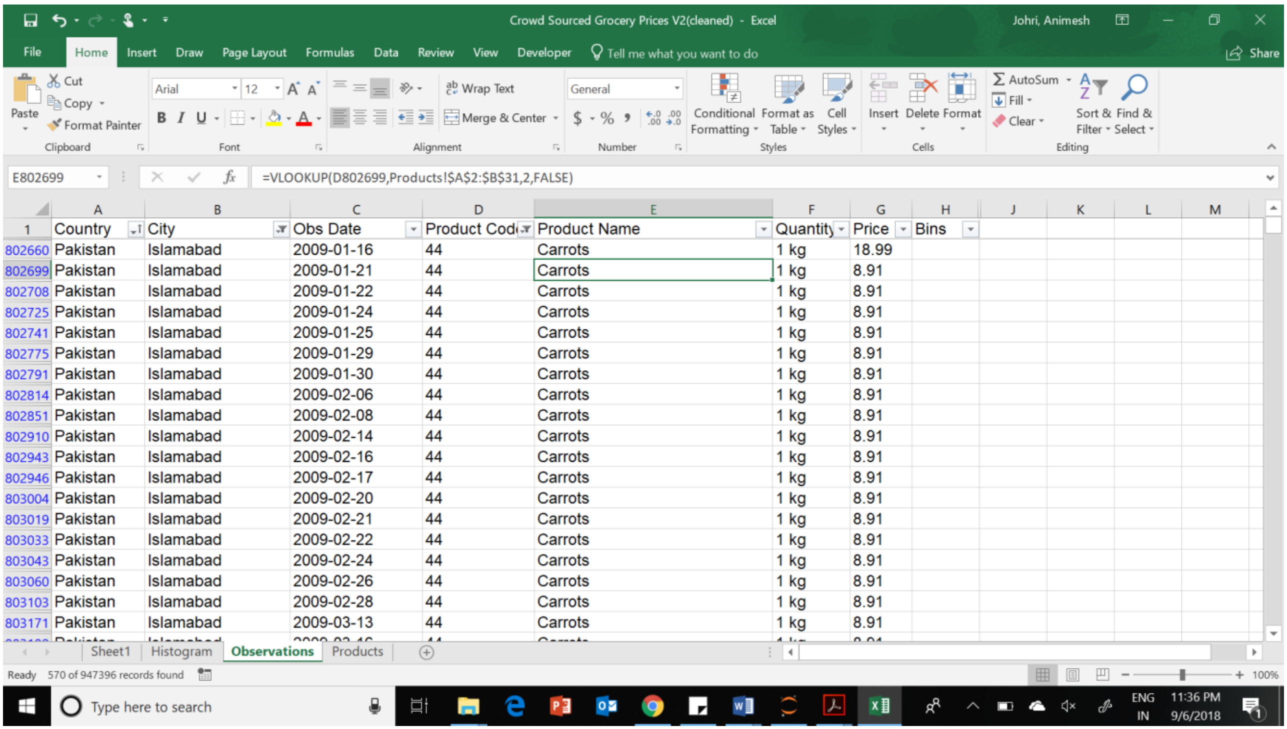
C1 | Obs Date

	A	B	C	D	E	F	G	H	J	K	L
1	Country	City	Obs Date	Product Cod	Product Name	Quantity	Price	Bins			
2	Kenya	Eldoret	2009-01-01	52	Cooking salt	1 kg	11.92	1		Number of	65
3	Pakistan	Attock	2009-01-01	44	Carrots	1 kg	9.36	2			
4	India	Nashik	2009-01-01	47	Potatoes, brown	1 kg	6.47	3			
5	Kenya	Eldoret	2009-01-01	52	Cooking salt	1 kg	5.91	4		Percentage	3.122384
6	India	Nashik	2009-01-01	62	Tea, black	100 g	1.80	5	0.376005		
7	Pakistan	Attock	2009-01-02	46	Maize	1 kg	16.97	6			
8	Kenya	Eldoret	2009-01-02	52	Cooking salt	1 kg	16.50	7			
9	India	Nashik	2009-01-02	47	Potatoes, brown	1 kg	10.43	8			
10	Kenya	Eldoret	2009-01-02	52	Cooking salt	1 kg	9.77	9			
11	Pakistan	Attock	2009-01-02	47	Potatoes, brown	1 kg	1.61	10			
12	Pakistan	Rawalpindi	2009-01-03	52	Cooking salt	1 kg	17.80	11			
13	Pakistan	Rawalpindi	2009-01-03	44	Carrots	1 kg	8.77	12		Median	8.91
14	Pakistan	Attock	2009-01-04	47	Potatoes, brown	1 kg	26.22	13			
15	Pakistan	Attock	2009-01-04	46	Maize	1 kg	20.07	14			
16	Pakistan	Islamabad	2009-01-04	45	Onion	1 kg	19.83	15			
17	Pakistan	Lahore	2009-01-04	44	Carrots	1 kg	19.35	16			
18	Pakistan	Islamabad	2009-01-04	47	Potatoes, brown	1 kg	15.86	17			
19	Pakistan	Islamabad	2009-01-04	46	Maize	1 kg	9.42	18			
20	Pakistan	Karachi	2009-01-04	46	Maize	1 kg	8.53	19			

Here the missing
more than 2%(3.1
for the price, ther
need to see the h
of the data. If hist
skewed then med
imputed if histogr
normal distributic
mean is imputed
missing blanks.

h. Product Name is “Carrots” and the Quantity of Product Code 44 in Islamabad is “1kg”.

Screenshot:



E802699 | =VLOOKUP(D802699,Products!$A$2:$B$31,2,FALSE)

	A	B	C	D	E	F	G	H
1	Country	City	Obs Date	Product Cod	Product Name	Quantity	Price	Bins
802660	Pakistan	Islamabad	2009-01-16	44	Carrots	1 kg	18.99	
802699	Pakistan	Islamabad	2009-01-21	44	Carrots	1 kg	8.91	
802708	Pakistan	Islamabad	2009-01-22	44	Carrots	1 kg	8.91	
802725	Pakistan	Islamabad	2009-01-24	44	Carrots	1 kg	8.91	
802741	Pakistan	Islamabad	2009-01-25	44	Carrots	1 kg	8.91	
802775	Pakistan	Islamabad	2009-01-29	44	Carrots	1 kg	8.91	
802791	Pakistan	Islamabad	2009-01-30	44	Carrots	1 kg	8.91	
802814	Pakistan	Islamabad	2009-02-06	44	Carrots	1 kg	8.91	
802851	Pakistan	Islamabad	2009-02-08	44	Carrots	1 kg	8.91	
802910	Pakistan	Islamabad	2009-02-14	44	Carrots	1 kg	8.91	
802943	Pakistan	Islamabad	2009-02-16	44	Carrots	1 kg	8.91	
802946	Pakistan	Islamabad	2009-02-17	44	Carrots	1 kg	8.91	
803004	Pakistan	Islamabad	2009-02-20	44	Carrots	1 kg	8.91	
803019	Pakistan	Islamabad	2009-02-21	44	Carrots	1 kg	8.91	
803033	Pakistan	Islamabad	2009-02-22	44	Carrots	1 kg	8.91	
803043	Pakistan	Islamabad	2009-02-24	44	Carrots	1 kg	8.91	
803060	Pakistan	Islamabad	2009-02-26	44	Carrots	1 kg	8.91	
803103	Pakistan	Islamabad	2009-02-28	44	Carrots	1 kg	8.91	
803171	Pakistan	Islamabad	2009-03-13	44	Carrots	1 kg	8.91	

Ready 570 of 947396 records found