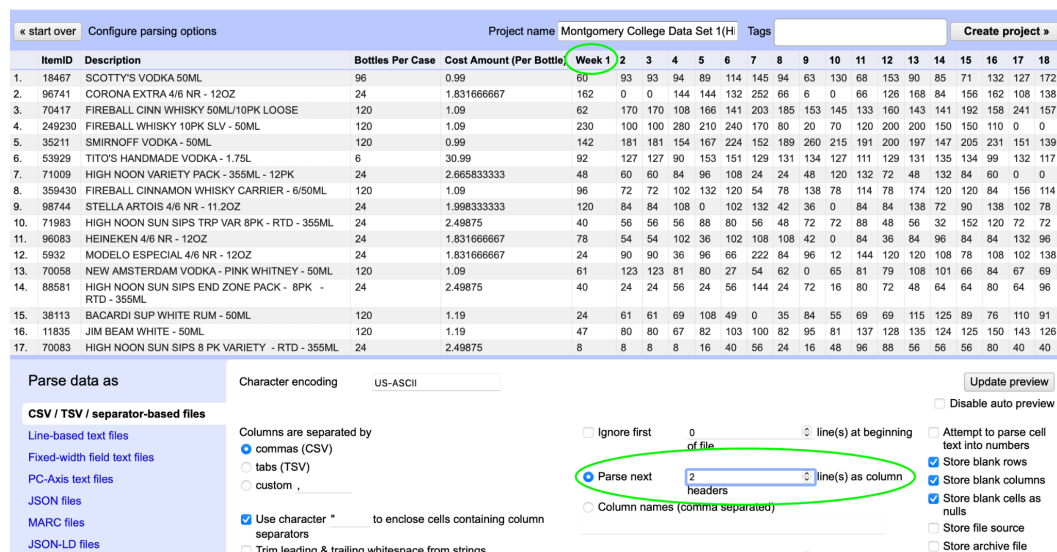


### Data Cleaning:

I will be working with **static data (CSV file)** using three data sets for this project to explore high-, medium-, and low-volume stores for inventory levels and algorithms.

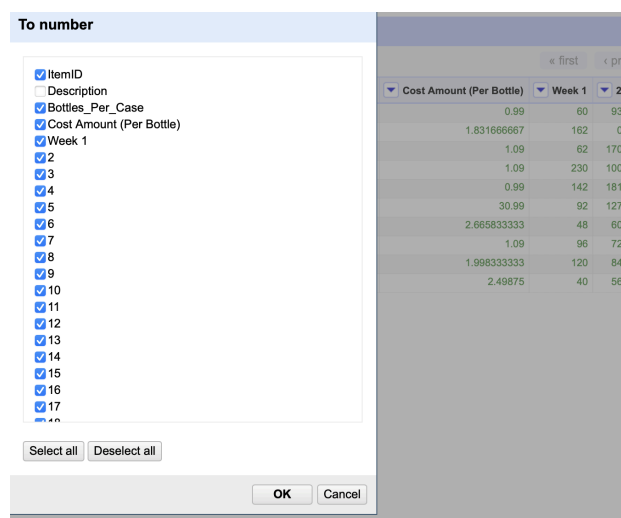
### Documentation:

First, I have to parse to the next line for the column name since Week was at the top. I will later have to change the name for each column name.



The screenshot shows the OpenRefine interface with a CSV file loaded. The 'Parse data as' panel is open, and the 'Parse next' option is selected for the column headers. The 'Column names (comma separated)' option is also selected. The 'Update preview' button is visible.

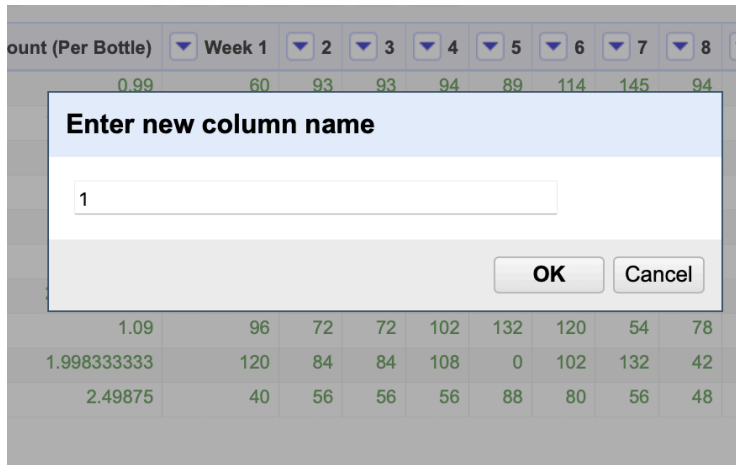
ItemID	Description	Bottles Per Case	Cost Amount (Per Bottle)	Week 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. 18467	SCOTT'S VODKA 50ML	96	0.99	60	93	93	94	89	114	145	94	63	130	68	153	90	85	71	132	127	172
2. 96741	CORONA EXTRA 4/6 NR - 12OZ	24	1.831666667	162	0	0	144	144	132	252	66	6	0	66	126	168	84	156	162	108	138
3. 70417	FIREBALL CINN WHISKY 50ML/10PK LOOSE	120	1.09	62	170	170	108	166	141	203	185	153	145	133	160	143	141	192	158	241	157
4. 249230	FIREBALL WHISKY 10PK SLV - 50ML	120	1.09	230	100	100	280	210	240	170	80	20	70	120	200	200	150	150	110	0	0
5. 35211	SMIRNOFF VODKA - 50ML	120	0.99	142	181	181	154	167	224	152	189	260	215	191	200	197	147	205	231	151	139
6. 53929	TITO'S HANDMADE VODKA - 1.75L	6	30.99	92	127	127	90	153	151	129	131	134	127	111	129	131	135	134	99	132	117
7. 71009	HIGH NOON VARIETY PACK - 355ML - 12PK	24	2.665833333	48	60	60	84	96	108	24	24	48	120	132	72	48	132	84	60	0	0
8. 359430	FIREBALL CINNAMON WHISKY CARRIER - 6/50ML	120	1.09	96	72	72	102	132	120	54	78	138	78	114	78	174	120	120	84	156	114
9. 98744	STELLA ARTOIS 4/6 NR - 11.2OZ	24	1.998333333	120	84	84	108	0	102	132	42	36	0	84	84	138	72	90	138	102	78
10. 71983	HIGH NOON SUN SIPS TRP VAR 8PK - RTD - 355ML	24	2.49875	40	56	56	56	88	80	56	48	72	72	88	48	56	32	152	120	72	72
11. 96083	HEINEKEN 4/6 NR - 12OZ	24	1.831666667	78	54	54	102	36	102	108	108	42	0	84	36	84	96	84	84	132	96
12. 5932	MODELO ESPECIAL 4/6 NR - 12OZ	24	1.831666667	24	90	90	36	96	66	222	84	96	12	144	120	120	108	78	108	102	138
13. 70058	NEW AMSTERDAM VODKA - PINK WHITNEY - 50ML	120	1.09	61	123	123	81	80	27	54	62	0	65	81	79	108	101	66	84	67	69
14. 88581	HIGH NOON SUN SIPS END ZONE PACK - 8PK - RTD - 355ML	24	2.49875	40	24	24	56	24	56	144	24	72	16	80	72	48	64	64	80	64	96
15. 38113	BACARDI SUP WHITE RUM - 50ML	120	1.19	24	61	61	69	108	49	0	35	84	55	69	69	115	125	89	76	110	91
16. 11835	JIM BEAM WHITE - 50ML	120	1.19	47	80	80	67	82	103	100	82	95	81	137	128	135	124	125	150	143	126
17. 70083	HIGH NOON SUN SIPS 8 PK VARIETY - RTD - 355ML	24	2.49875	8	8	8	8	16	40	56	24	16	48	96	88	56	56	80	40	40	40



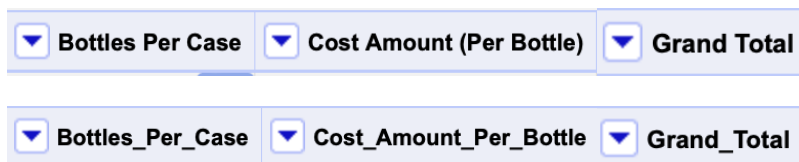
The screenshot shows the 'To number' dialog box in OpenRefine. The 'ItemID' column is selected, and the 'Week 1' column is being converted to numerical data. The 'OK' button is visible.

ItemID	Description	Bottles Per Case	Cost Amount (Per Bottle)	Week 1	2
1. 18467	SCOTT'S VODKA 50ML	96	0.99	60	93
2. 96741	CORONA EXTRA 4/6 NR - 12OZ	24	1.831666667	162	0
3. 70417	FIREBALL CINN WHISKY 50ML/10PK LOOSE	120	1.09	62	170
4. 249230	FIREBALL WHISKY 10PK SLV - 50ML	120	1.09	230	100
5. 35211	SMIRNOFF VODKA - 50ML	120	0.99	142	181
6. 53929	TITO'S HANDMADE VODKA - 1.75L	6	30.99	92	127
7. 71009	HIGH NOON VARIETY PACK - 355ML - 12PK	24	2.665833333	48	60
8. 359430	FIREBALL CINNAMON WHISKY CARRIER - 6/50ML	120	1.09	96	72
9. 98744	STELLA ARTOIS 4/6 NR - 11.2OZ	24	1.998333333	120	84
10. 71983	HIGH NOON SUN SIPS TRP VAR 8PK - RTD - 355ML	24	2.49875	40	56

Here, I converted the data that is numeral to numerical data, since it was first labeled as strings.



For each data set, I changed the column (Week 1) to “1” that way it’s readable and I can pull out each week with the data as I am working only defining the number of the week.



Changed Space in Variables with an underscore “\_”.

Finally, I will use R Studio and save these data sets to make a new column, because I want to take the...

“Bottles\_Per\_Case” and **MULTIPLY** it by “Cost\_Amount\_Per\_Bottle” to get the total cost of the case of alcohol in the stores.

You can open up the code file on R Studio where I have done the below:

```

Finally, I will do my last step with this data set, which is to multiply to "Bottles_Per_Case" variable with
"Cost_Amount_Per_Bottle" variable to get the "Total_Cost" variable for all the bottles in a case. I will do this for all three data
sets.

```{r}
# Creating the Total_Cost variable
High_Volume_Weekly_DS <- High_Volume_Weekly_DS |>
  mutate(Total_Cost = Bottles_Per_Case * Cost_Amount_Per_Bottle)

Medium_Volume_Weekly_DS <- Medium_Volume_Weekly_DS |>
  mutate(Total_Cost = Bottles_Per_Case * Cost_Amount_Per_Bottle)

Low_Volume_Weekly_DS <- Low_Volume_Weekly_DS |>
  mutate(Total_Cost = Bottles_Per_Case * Cost_Amount_Per_Bottle)

```

```

I also decided to change the number columns for weeks as “Week #”

```

I noticed that the variables on the top of my data set have X1, X2, X3, etc. I will remove the X and replace it with "Week".

```{r}
# Renaming the columns
colnames(High_Volume_Weekly_DS) <- gsub("X", "Week", colnames(High_Volume_Weekly_DS))
colnames(Medium_Volume_Weekly_DS) <- gsub("X", "Week", colnames(Medium_Volume_Weekly_DS))
colnames(Low_Volume_Weekly_DS) <- gsub("X", "Week", colnames(Low_Volume_Weekly_DS))

```

```

The data is officially cleaned!

Now onto my EDA! I will continue to work on this on my RStudio Markdown.