Process of Vehicle Count to Company Distribution Analysis (Filtered and Non-Filtered)

1. First, I filtered the datasets based on whether the year is over/under 2008.
   Then, I extracted all of the unique companies for those sets.
   Last for the first step, for each set, I was able to aggregate the sum of each record for the unique companies.

2. Even though this was a little extraneous, I created percentage breakdowns of 65, 75, 85, 95% groupings of all trucks in the whole set.
   I created a cumulative sum of the trucks and a cumulative sum of the total percentage of the vehicles in each set.
   I looked at whether there were any "big shifts" around these percentages to see if there was anything necessary to note, but there wasn't so, I broke off the 75% of records and then top 25% of records (or close cut-offs). This showed the companies that have the largest truck counts that represent 25% of the trucks.
   It was noted how many trucks per company represented the 75% cutoff.

3. It was then determined that there weren't enough sensible "shifts" away from the major quadrants and it made more statistical sense to break these sets down near the quarter cut offs with regard to the total vehicle population. It was easiest to build a function to look at records near the 25, and 50% cut offs and see if there were any big "jumps" or changes in trend **near those**, and if not, just extract the closest percentage cut offs at those. A function was built to see the closest 10 records near these different percentage cut offs. The associated percentages were noted.
   In this, another column was created that showed the total amount of vehicles (in case more than one company had the same amount of filtered/non-filtered vehicles) in a particular "Truck Count" group. (Note it was assumed early that these vehicles were all trucks so Trucks represent vehicles).

Now that the appropriate 25%, 50% and 75% cut offs are in place, labels were created in a function, and a new column was added to each dataset showing which "quadrant" each company fit in based on the amount of trucks (filtered or not) were on the road.
Here, now that the groups were all separated out, there was now a total number of groups in each quartile.

Here is an example:

Quartile	Company count	Cumulative %
1st quarter (> 96 trucks)	121	0.252195
2nd quarter (96>= truck count > 27)	514	0.502778
3rd quarter (27>= truck count >=9)	1597	0.738891
4th quarter (9 > truck count)	9683	1.000000

QA Checks: Next, a column was created to see the number of companies in each Truck count group. We checked that the original counts of companies matched the total amount of companies in our company counts once it has been separated. That checked out (*They would be false if the values weren't equal*):

```
[43]  1 #here is a QA check to make sure the company counts are the same
      2 org_cont = non_filt['Owner Name'].unique()
      3 i=org_cont.size
      4 cnt_sum =counts_nf['Company count'].sum()
      5
      6 org_cont_filt = filtered['Owner Name'].unique()
      7 j = org_cont_filt.size
      8 cnt_sum_filt = counts_f['Company count'].sum()
      9
     10 i==cnt_sum, j== cnt_sum_filt

     (True, True)
```

Then we checked that the percentage breakdowns were close to the actual breakdowns (considering that the actual breakdowns didn't hit exactly on the exact 25% divider marks). These numbers were very close and therefore checked out as well indicating the process

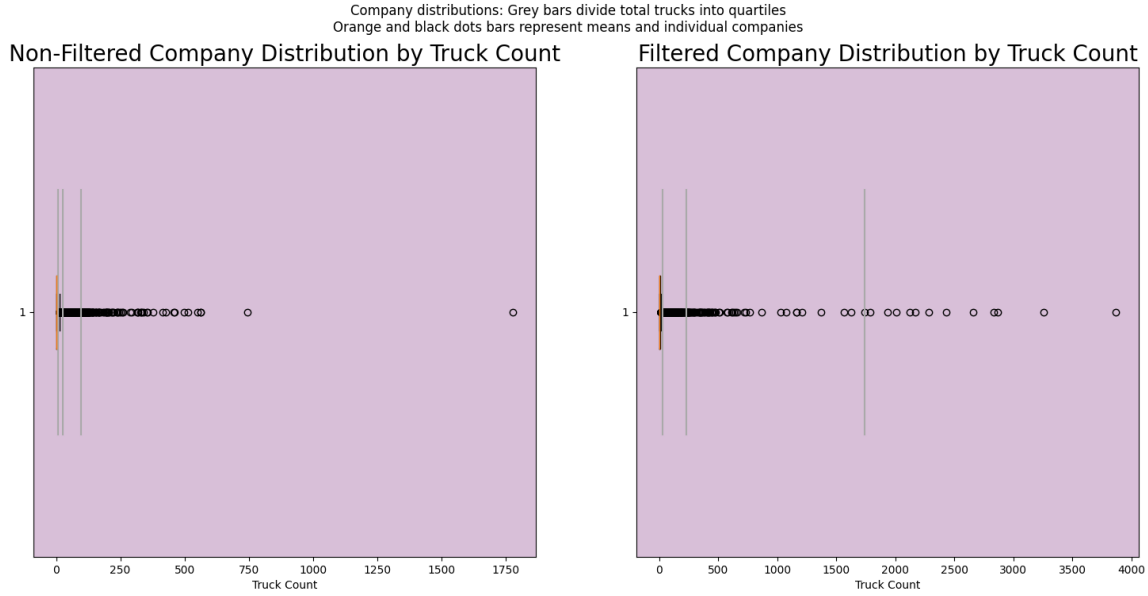
was accurate (within a very, very, high probability):

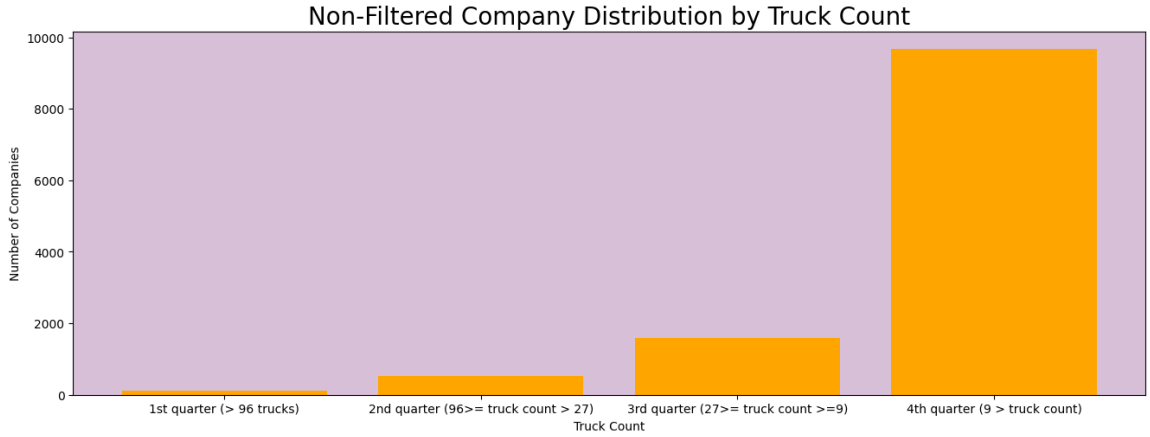
Percent	Quarters Non-Filt	df_nf Quartile	Quarters Filt	df_f Quartile
0.25	24654.5	24871	32945.0	31959
0.5	49309.0	49583	65890.0	65787
0.75	73963.5	72868	98835.0	99319

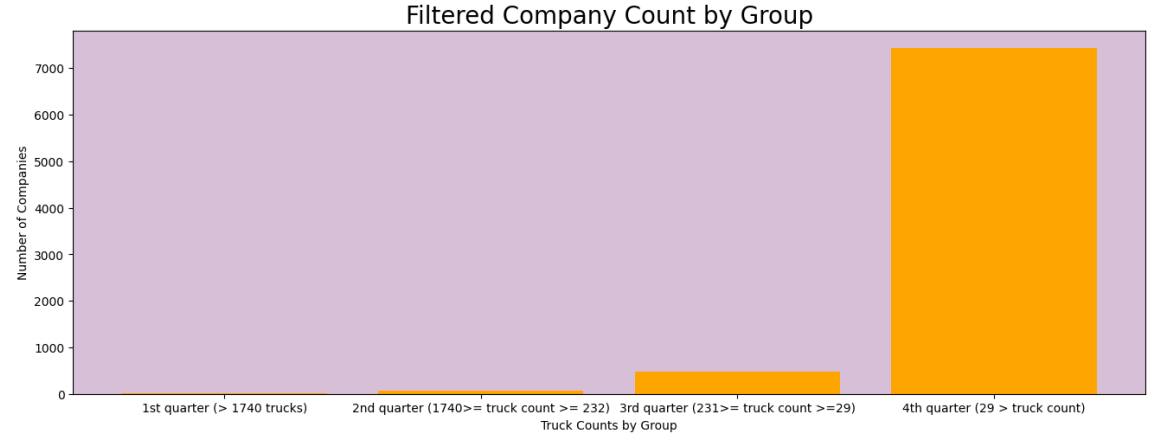
5. Ok, finally, I looked at the overall spread of the companies based on the number of trucks they had. Then, I created a chart that showed the quadrant cut offs based on truck counts, as well as the box-plot values (mean, quadrants) based on just the companies and how many trucks they have. So, the grey bars show total sum of truck quadrants, and the orange bars show the median of companies that are skewed as you can see. This means there are many more companies with small truck counts than large truck counts.

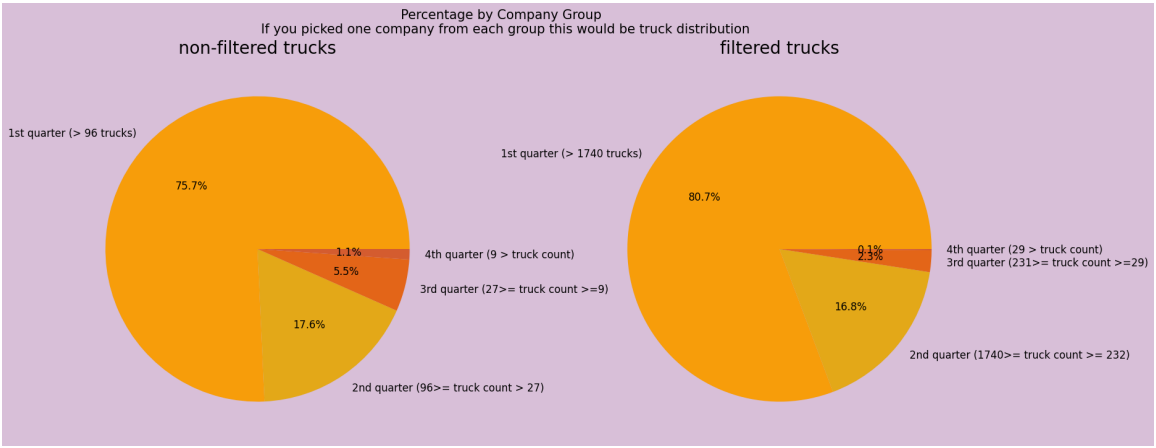
6. Finally, I graphed the number of companies in each category (the more trucks a company has the less companies will be in its category). I also thought a pie chart would exemplify how much each sized company might contribute to the total amount of trucks on the road. I created a filtered and non-filtered graph for all of this. You can see all of this in the charts below.

Here are the charts for the vehicles which are over 26000 lbs. from ODOT:





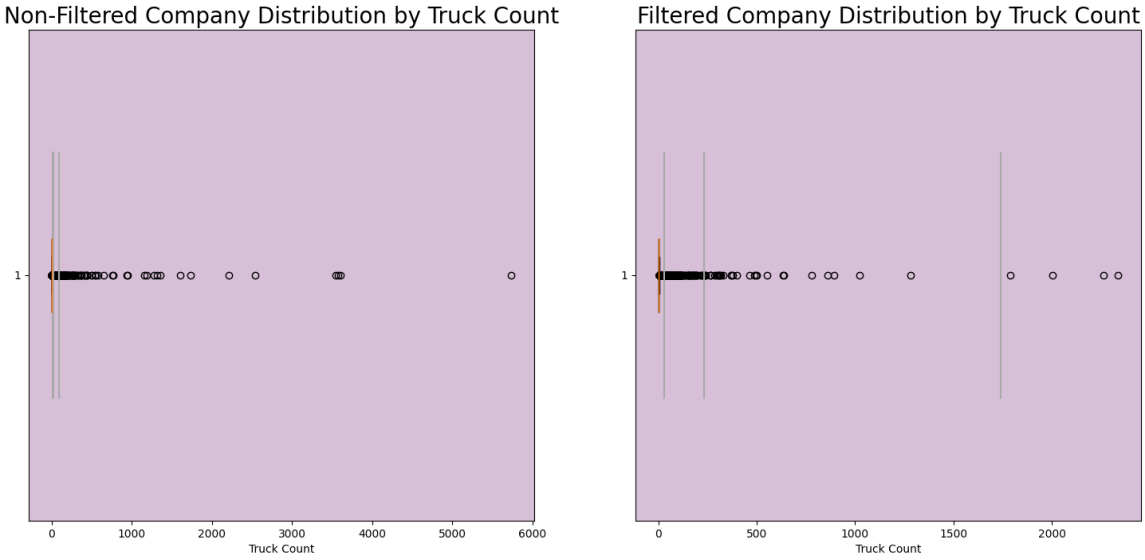


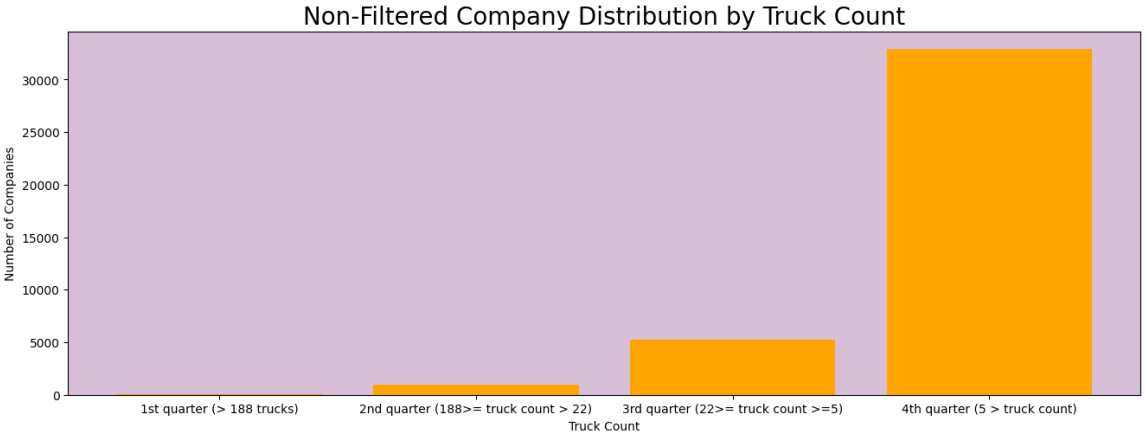


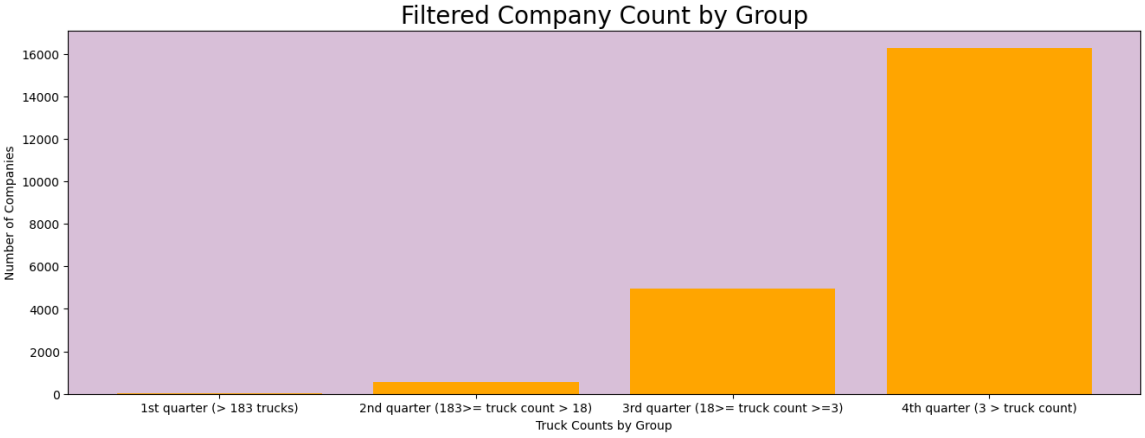
It is hard to save these in a document so there were PNG files saved for each of these.

Here is the breakdown for vehicles under 26000 pounds from DMV:

Company distributions: Grey bars divide total trucks into quartiles
Orange and black dots bars represent means and individual companies







Note this last chart was rotated because it was a little easier to read the labels.


Percentage by Company Group
If you picked one company from each group this would be truck distribution
non-filtered trucks
4th quarter (5 > truck count)
3rd quarter (22>= truck count >=5)
2nd quarter (188>= truck count > 22)
1st quarter (> 188 trucks)
0.2%
1.0%
5.7%
93.1%
filtered trucks
4th quarter (3 > truck count)
3rd quarter (18>= truck count >=3)
2nd quarter (183>= truck count > 18)
1st quarter (> 183 trucks)
0.2%
1.0%
7.2%
91.6%