Sample Data Analysis Steven Gaetz

Notes on errors

The first thing I did was open the data set sample in Excel. I sorted the first column, Unit Number, from lowest to highest. This resulted in four null values in this column at the foot of the data set. It did not make sense to me that a company wouldn't have numbers for any of their units. Looking across these four rows through the rest of the columns, it appeared that every value has been shifted to the right by one column or that someone put the data into the spread sheet wrong. I highlighted these rows, copied, and pasted the values to the left by one column. All the values made much more sense when compared to the rest of the data. The only column that looked off now was the Life to Data Mileage column, which had values that looked more like a VIN or other code. I would keep this in mind if using this variable for analysis.

I then put the data set into Tableau and made a few quick bar graphs, looking at each Unit Number in rows and different variables into columns, such as Annual Mileage, Fuel Cost, and Fuel Gallons. After sorting in descending order, I noticed that unit 18263 was showing up with some crazy high values that did not make sense. I also noticed that the number of rows in Tableau were 137 less than the number of rows that there were in the Excel file. I thought that there must be some duplicates, so I highlighted duplicates in column A and indeed there were 137 duplicates of unit 18263. I went ahead and deleted the extra rows so that now there was only one row of this unit.

Exploring more of the columns in the dataset, I noticed there was no difference between the Purchase Price Without Tax and the Purchase Price with Tax, except when I sorted the latter from largest to smallest, I found the error. There are three extra zeros on the Price w/ Tax for units 13319 and 25184. I removed the extra zeros and now the prices matched up.

Another column I noticed an error in was License Cost. When sorted from largest to smallest value, the largest was extremely large compared to the other values. This row was unit 96243, which is a 1996 GM/Chevy 1500 pickup truck. With a little digging I found that the same truck near that year had license costs of $176.83, so I removed the extra zeros in the suspect cell and made the cost this number.

The next error I found was in the column for Mechanic Labor Cost. I found this the same way, I sorting the column from largest to smallest value and compared the largest value to the rest of the values in the column and it clearly did not make sense. Mechanic Labor should not cost $133800000.44, and I figured that someone again mistakenly put zeros into this value. I removed the extra zeros making the cost $1338.44. This is much more reasonable.

The next error was similar. It was in the Parts Cost column. I proceeded the same as the last error, compared the value to similar vehicle, and removed zeroes to end up with a cost of $376.32.

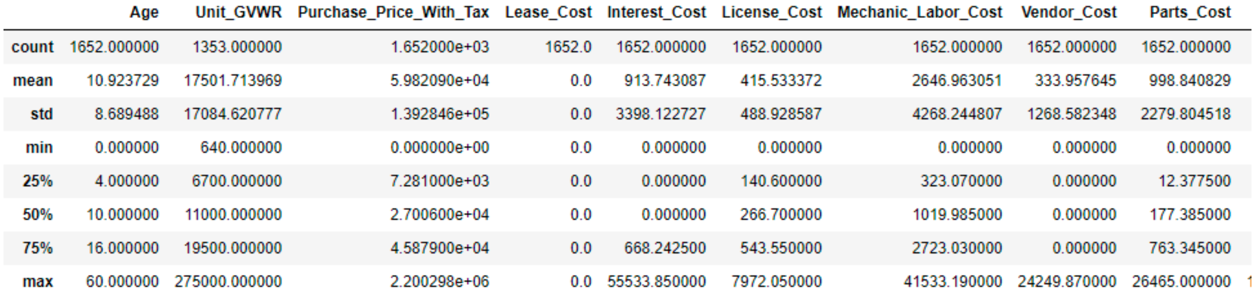
Working in the same fashion across the columns, the next error I found was in the Total Cost column. The highest value was $2,468,499 for unit number 10408. This unit had a purchase price of $108,461, so the total cost for this unit does not make sense. I went through all the costs for this row and they added up to $41,141.65. I put this value into the cell.

The column of Unit GVWR (gross vehicle weight rating) had a couple of extreme values, 275,000 and 110,000 for units 11867 and 11503, respectively. These seemed a bit too high, compared to the rest of the data in this column. Similar trailers did not have values for GVWR, so I left the values alone since I was not 100% sure that they weren't correct.

I notice one more thing that I found odd. In Tableau I created a new column Miles Per Gallon. Unit 29656 had the best 679 miles per gallon! Its annual mileage was 3395 and fuel gallons was 5. This is a Bobcat excavator, which when in use is probably stationary. These numbers are probably wrong, but miles per gallon is probably not the way to measure fuel economy in this type of vehicle. I would think it would be hours of run time per gallon of gas or something similar. I decided to leave it the way it is, since it is in the Utilimarc Category of POE. Fuel economy might make more sense in the Vehicle Category.

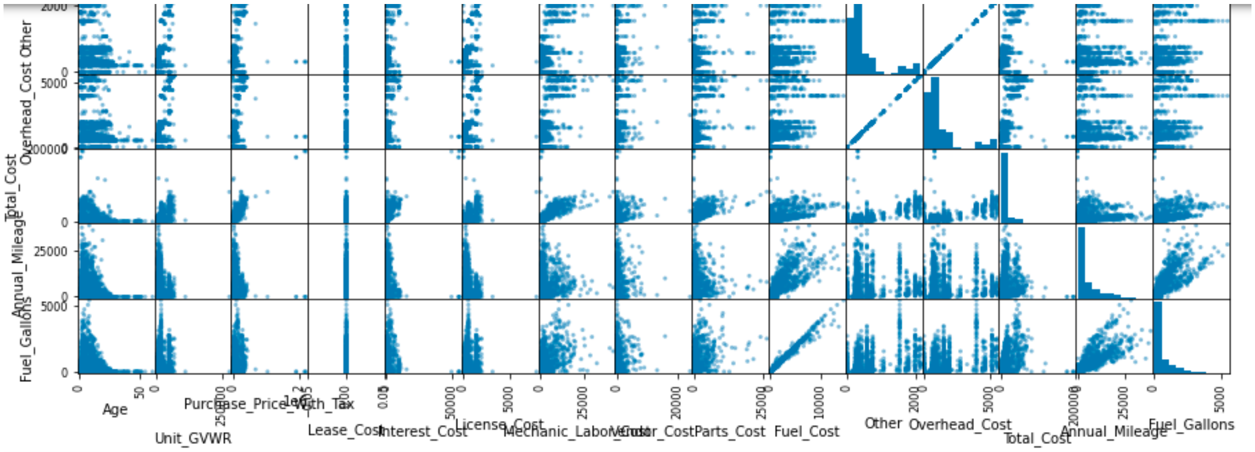
Analysis

First, I created a new repository in GitHub so I had a place to save my work. I cloned the repository on my computer and saved the cleaned data in this directory. I used Python in a Jupyter Notebook. I imported the cleaned data, as well as pandas and pyplot from matplotlib. I created a subset of the data because I new there were some variables, such as Status, Description, etc., that I was did not want to deal with since my time was limited. I printed out some descriptive statistics using so I could quickly reference the mean, max, and min for each variable.



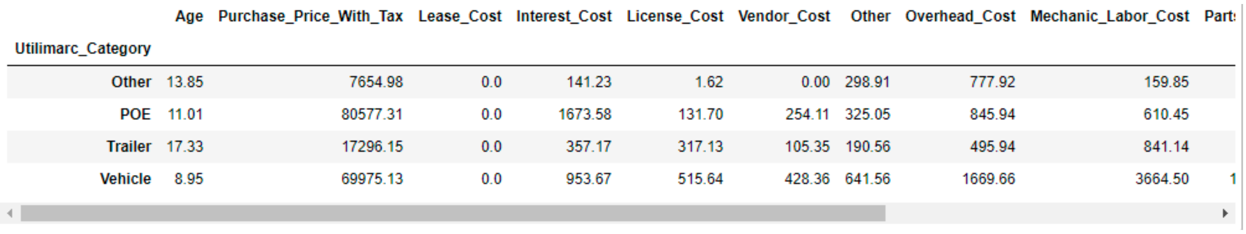
	Age	Unit_GVWR	Purchase_Price_With_Tax	Lease_Cost	Interest_Cost	License_Cost	Mechanic_Labor_Cost	Vendor_Cost	Parts_Cost	
count	1652.000000	1353.000000	1.652000e+03	1652.0	1652.000000	1652.000000	1652.000000	1652.000000	1652.000000	
mean	10.923729	17501.713969	5.982090e+04	0.0	913.743087	415.533372	2646.963051	333.957645	998.840829	
std	8.689488	17084.620777	1.392846e+05	0.0	3398.122727	488.928587	4268.244807	1268.582348	2279.804518	
min	0.000000	640.000000	0.000000e+00	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	4.000000	6700.000000	7.281000e+03	0.0	0.000000	140.600000	323.070000	0.000000	12.377500	
50%	10.000000	11000.000000	2.700600e+04	0.0	0.000000	266.700000	1019.985000	0.000000	177.385000	
75%	16.000000	19500.000000	4.587900e+04	0.0	668.242500	543.550000	2723.030000	0.000000	763.345000	
max	60.000000	275000.000000	2.200298e+06	0.0	55533.850000	7972.050000	41533.190000	24249.870000	26465.000000	1

I also created a scatterplot matrix to get a quick view of each variable plotted against one another to see if any thing jumped out at me.



It was no surprise that Fuel Cost and Fuel Gallons are strongly correlated or that Other and Overhead cost are as well.

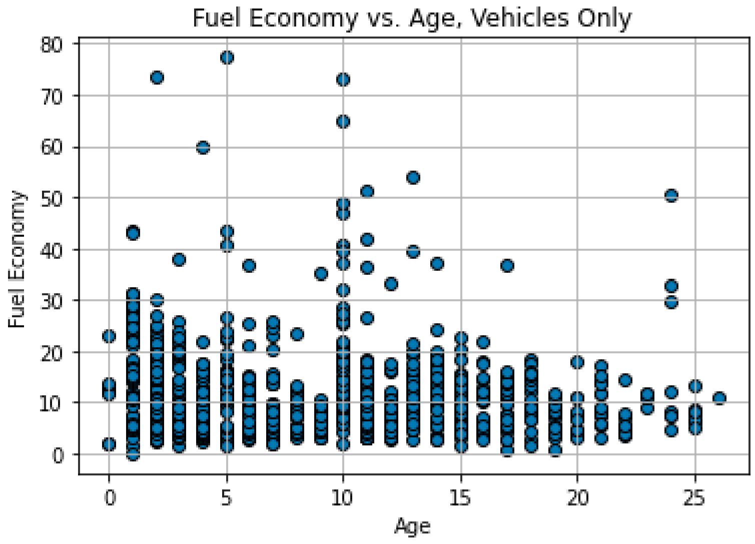
I decided to group the data by Utilimarc Category because I was really interested in comparing some numbers between Vehicles, Trailers, POE, and Other.



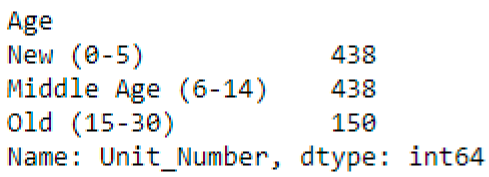
	Age	Purchase_Price_With_Tax	Lease_Cost	Interest_Cost	License_Cost	Vendor_Cost	Other	Overhead_Cost	Mechanic_Labor_Cost	Part:
Utilimarc_Category										
Other	13.85	7654.98	0.0	141.23	1.62	0.00	298.91	777.92	159.85	
POE	11.01	80577.31	0.0	1673.58	131.70	254.11	325.05	845.94	610.45	
Trailer	17.33	17296.15	0.0	357.17	317.13	105.35	190.56	495.94	841.14	
Vehicle	8.95	69975.13	0.0	953.67	515.64	428.36	641.56	1669.66	3664.50	1

Now I could see that the average price for a POE is $80577.31, more than $10,000 more than the average price of a vehicle. The average cost of a Vehicle License is $515.64, higher than the other categories. Vehicles was higher in every cost except for interest, which was highest for POE since its price was the highest. Vehicles make up the bulk of Annual Mileage, which I had suspected when I was cleaning the data.

I decided to explore Vehicles on their own since they had such high costs and I wanted to look at some fuel economy numbers. I created a Fuel_Econ variable, first dropping the zeros from Annual Mileage and Fuel Gallons. I created a plot of Fuel Economy versus Age because I wanted to see if there was any real difference as age increased.

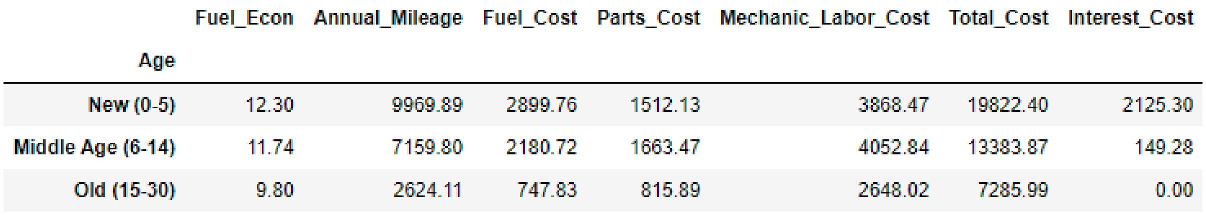


It is not crystal clear, but I'd say there is a slight drop in fuel economy, as you might expect from a newer vehicle versus an older one. I decided to put the Age variable into three bins and see if there was much difference between three different age groups. Vehicles that were 5 years old or less I labeled "new", 6 to 14 years old I labeled "middle age", and 15 years and older were "old". The number of each came out as follows:



```
Age
New (0-5)           438
Middle Age (6-14)   438
Old (15-30)         150
Name: Unit_Number, dtype: int64
```

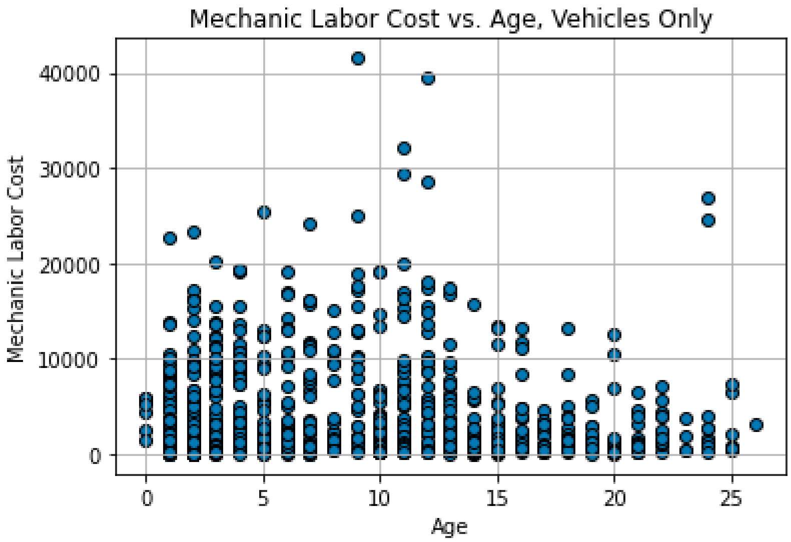
I think these bins could be played around with a little, but I thought this was good enough for now. I made a new table with some averages to look at.



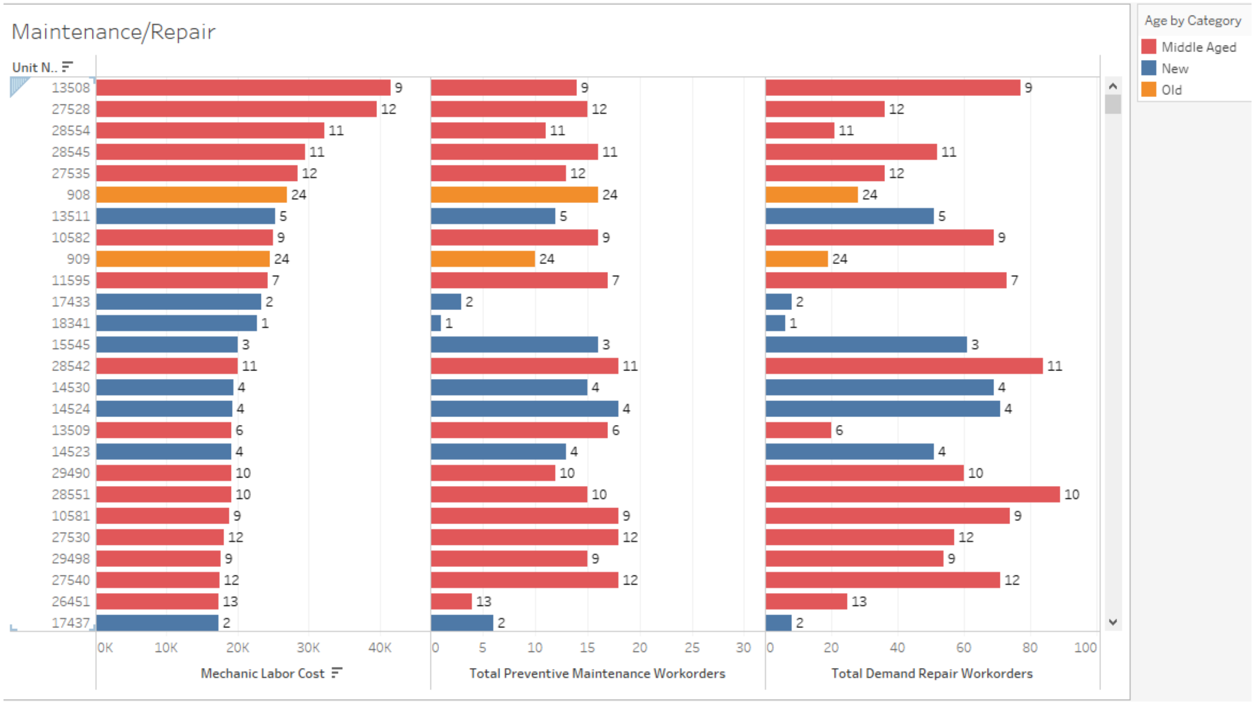
	Fuel_Econ	Annual_Mileage	Fuel_Cost	Parts_Cost	Mechanic_Labor_Cost	Total_Cost	Interest_Cost
Age							
New (0-5)	12.30	9969.89	2899.76	1512.13	3868.47	19822.40	2125.30
Middle Age (6-14)	11.74	7159.80	2180.72	1663.47	4052.84	13383.87	149.28
Old (15-30)	9.80	2624.11	747.83	815.89	2648.02	7285.99	0.00

New cars on average were 2.5 mpg more efficient than old cars. They also averaged over 7,000 miles more per year and obviously had a much higher fuel cost because of this. I also noticed some differences in parts cost and mechanic labor cost between the three groups. This is what I explored next.

Here is a scatterplot of Mechanic Labor Cost versus Age:

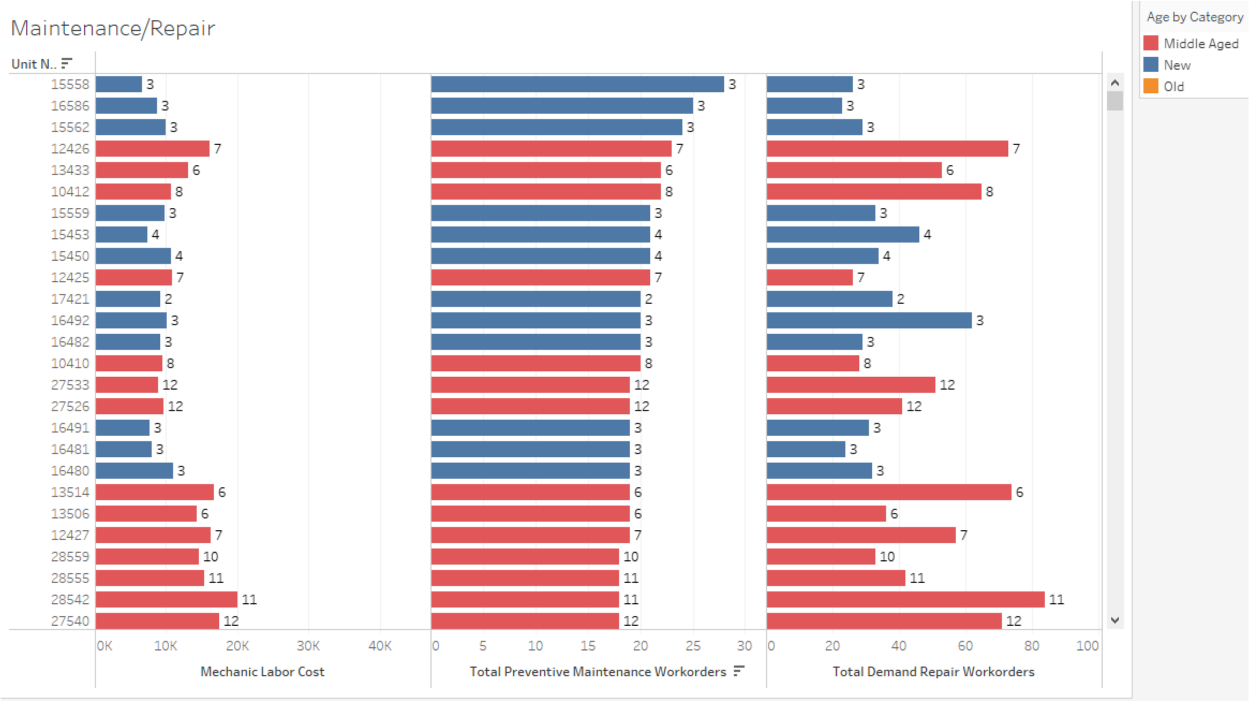


It appeared that mechanic labor cost increases to middle age and then decreases as vehicles get old. This makes sense since newer vehicles should not need as much work. There are also less older vehicles, and maybe older vehicles are cheaper to work on. I decided to go back to some of the visualizations that I had been playing around with earlier.



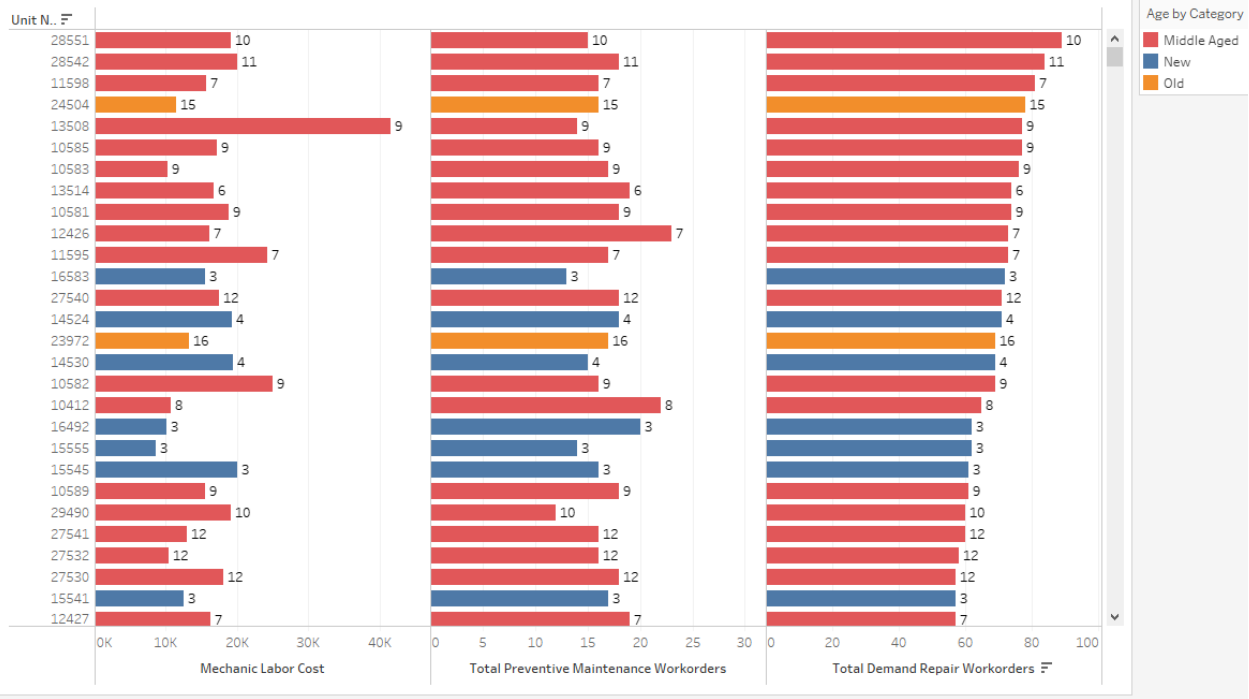
Here is a look at the Units with the highest Mechanic Labor Cost, as well as their total preventive maintenance workorders and their total demand repair workorders. The age of the vehicle is shown at the end of the bar. Vehicles that are middle aged are clearly dominated here.

I resorted by the highest number of Total Preventative Maintenance Workorder:



As we would expect, newer cars came to the top of the list. Although they have a high number of workorders, they have a much lower labor cost and less Repair Workorders.

I resorted for the Total Demand Repair Workorders:



Vehicles that are middle aged again rise to the top of the list. I think the smaller number of old vehicles is the reason they don't show up in these bar charts a whole lot. Again, I think that the age ranges could be tweaked a bit too.