THE UNIVERSITY of NORTH CAROLINA
GREENSBORO

2017

Predictive Analytics Solution: Old Dominion Gas Mileage

Orrin Wheeler
University of North Carolina Greensboro
11/10/2017

Contents

Predictive Analytics Solution – Linehaul and PND MPG Exploration and Predictions

Old Dominion Freight Line (ODFL) is one of the premier less than truckload (LTL) shipping companies in the continental United States. With over 10 million shipments in 2016, ODFL delivers goods to thousands of customers daily. As with all companies, we want to keep our operating costs low to maximize capital available for investments in the company's future and to better suit our customers needs. One of ODFL's biggest costs is keeping fuel in the tanks of the fleet. Since ODFL tractors moved over 500 million miles in 2016 alone, employing drivers and fielding trucks that can consume fuel as efficiently as possible can save the company hundreds of thousands of dollars each year.

The routes ran by Old Dominion come in two varieties, pickup and delivery (PND) and linehaul. These two routes are very different in their identifiable characteristics. Linehaul routes involve over the road trucking, moving freight from one service center to another. Linehaul routes involve higher speeds and less downtime than their stop and start sister PND. It is hypothesized that maximizing the time spent using the cruise control and the topmost gear will produce the highest MPG ratings.

PND routes involve either taking cargo from a service center to a consumer and dropping it off, or conversely picking up cargo from a consumer and bringing it back to a service center to be shipped across the country. PND routes tend to have more starts/stops and involve more idle time than linehaul. Thus we expect that minimizing both long and short idle time as well as brake events will result in the best MPG rating. It is also expected that for both route types, more tenured drivers will produce a better MPG rating.

The goal of this project is to develop analytics based solutions to identify traits and characteristics of routes that produce a high miles per gallon ratio. A deep exploration into the data resulting from thousands of data points taken directly from sensors on the tractors themselves will be conducted using descriptive and predictive modelling in the hopes of achieving that goal. There will be two models produced, one predicting the MPG of linehaul routes, and one predicting the MPG of PND routes. These models can be used in maximizing driver efficiency as well as informing future studies into minimizing the costs incurred by fuel consumption among the fleet.

The data source

The dataset being used for this study was pulled from multiple sources. The initial set contained quantitative data from sensors located on tractors throughout the ODFL fleet while on routes run during the month of June 2017. This dataset contained 66 variables and 570,911 individual observations. This was then combined with qualitative information taken from the data service center. This qualitative information described both the tractors and the drivers associated with each dispatch. Since we are attempting to predict the MPG for a given route, the MPG variable was calculated from the packet data for use as the target variable.

Once these sets were combined, we then began cleaning in order to reach a finalized set on which we could perform proper analytics. An in depth explanation of all variables involved in this project can be found in Appendix A.

Data cleansing and transformations

First, tractor packet observations are grouped by the tractor and driver combination used on each dispatch. Then, to separate each dispatch, packet data observations are aggregated by combining all observations that begin within the starting and end points indicated by the dispatch table. By doing this, packets that contain information about tractor movements outside of dispatches (i.e. moving a tractor across a yard or moving a trailer from parked location to the terminal to be loaded) were removed so as to not affect the analysis to be done.

To remove bias based on the length of a packet, a variable was computed to contain the percent of miles within a packet each driver was using either the top gear of the tractor or the cruise control. Doing this helped to normalize the information about the top gear distance and cruise distance.

Additionally, any data point in which the travel miles was less than two, or the miles per gallon was greater than 20 are removed. These values were identified as outliers and would negatively impact the accuracy of any model produced. Lastly, the PACKET_MILES variable was dropped from the dataset because it was determined that while this may have a significant impact on the MPG, ultimately the number of miles travelled in a given dispatch is out of the control of the driver and thus should not count for nor against them when predicting their MPG. These transformations resulted in a dataset containing 35 variables and 15,062 data points.

Descriptive Statistics

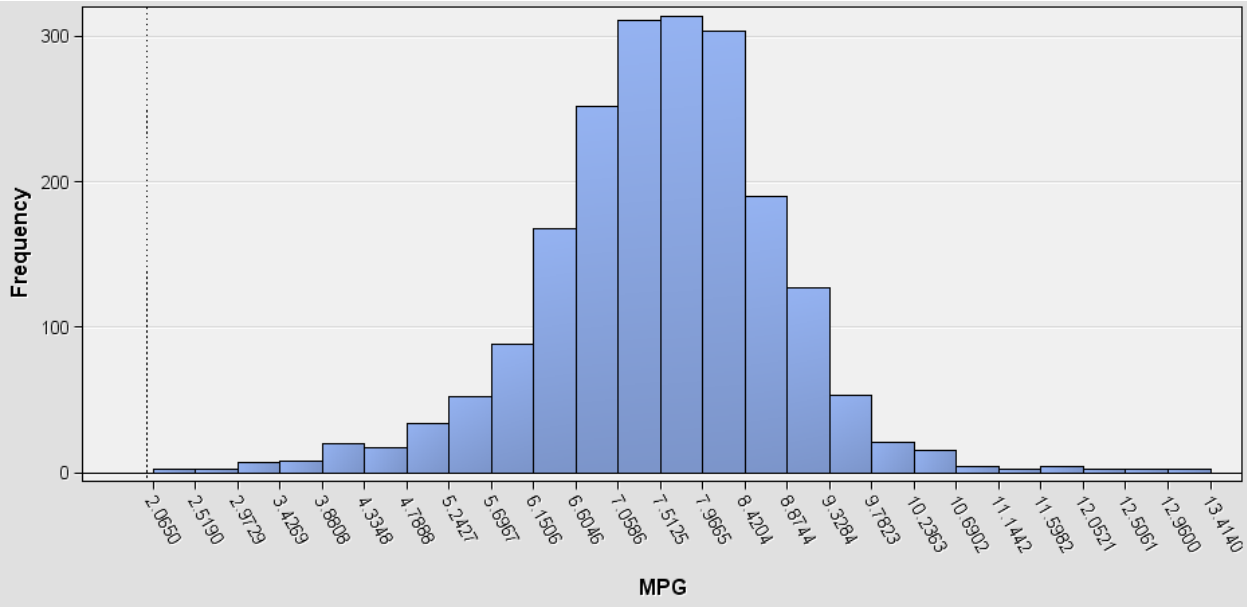
Descriptive analysis is performed to summarize characteristics and identify patterns within the dataset.

Exploring and validating the contents of the data is a vital step in the validation of any final product of analysis. This validation helps to ensure accuracy, observe possible patterns, and minimize risk of errors within the analysis.

Using the Stat Explore node to return a series of summary statistics, we are given clarity into the nature of the data. For interval variables, values returned include the mean, standard deviation, the number of missing values, the minimum, median, maximum, skewness and kurtosis of each variable. For nominal variables, the number of levels for each variable is presented along with the first and second modes along with their percentage of appearance within the data set.

The MPG Variable

Before beginning to produce a model to predict the MPG of a given data packet, an exploration of the target MPG variable is necessary. As a whole, the MPG variable produces a very nice normal curve as seen in the graph below. It has a skewness of only .02 and kurtosis of 3.11. These values indicate a near-normal distribution.



1 MPG distribution

This normally distributed target variable allows for more accurate confidence intervals during the analysis phase.

PND vs Linehaul separation

One final alteration was performed on the dataset where the data was split into two partitions. The first dataset contains data pertaining to dispatches identified as linehaul routes by the DISPATCH_TYPE variable, while the other contains data pertaining to dispatches identified as PND routes. We determined that the independent nature of these routes made the split necessary as each partition will contain different indicators of good gas mileage. This allows for a more representative dataset for each type of route run by Old Dominion Freight Line.

This separation of data results in a Linehaul dataset with 8,275 observations and a PND dataset with 6,817 observations. Both sets still contain > 5,000 observations so proper statistical analysis can still be performed.

Linehaul Descriptive Analysis

Following is a descriptive analysis of the linehaul partition and a presentation of observations to be made from it.

Variable Name	Levels	Missing	Mode	Mode Percentage	Mode2	Mode 2 Percentage
REP_CAB_TYPE	3	0	DAYCAB	91.7	SLEEPER	7.92
REP_DRIVE_AXLE_SET_UP	2	0	SINGLE AXLE	96.26	TANDEM AXLE	3.74
REP_ELOG_CERTIFIED_FLG	3	0	Y	97.84	N	2.09
REP_ELOG_TRAINED_FLG	3	0	Y	97.84	N	2.09
REP_ENGINE_MAKE_TXT	4	0	DETROIT	69.51	CUMMINS	26.39
REP_ENGINE_MODEL	9	0	DD15	65.22	ISX	25.52
REP_EQUIPMENT_CATEGORY_TYPE_NM	4	0	DAYCAB	90.72	SLEEPER	7.92
REP_FIFTH_WHEEL_TYPE	2	0	FIXED	97.11	SLIDE	2.89
REP_FULL_TM_FLG	2	0	F	99.92	P	0.08
REP_LINEHAUL_PACOS_TRAINED	3	0	Y	97.84	N	2.09
REP_MAKE	2	0	FRGHT	92.87	VLVNA	7.13
REP_MODEL	7	0	CA125	87.47	VNL42T300	6.68
REP_MODEL_YEAR	14	0	2016	28.13	2017	21.15
REP_PND_PACOS_TRAINED	2	0	N	99.93	U	0.07
REP_REAR_AXLE_RATIO	8	0	3.42	93.52	3.58	4.44
REP_TRANSMISSION_MAKE	4	0	EATON	95.66	MERTR	2.67

2 Linehaul Nominal Variables

In viewing the output for the nominal linehaul variables, a few observations can be made.

1. Only 4 variables have a mode that appears in less than 90% of the data (REP_ENGINE_MAKE_TXT, REP_ENGINE_MODEL, REP_MODEL, REP_MODEL_YEAR)
2. 91.7% of linehaul routes are using day cabs
3. The most common rear axle ratio is 3.42 with 93.5% of dispatches
4. REP_MODEL_YEAR is most commonly 2016 at 28.1% and the second most common year being 2017.
5. 99.9% of drivers are full time and have received the PND PACOS Training while 97% have received the Linehaul PACOS training

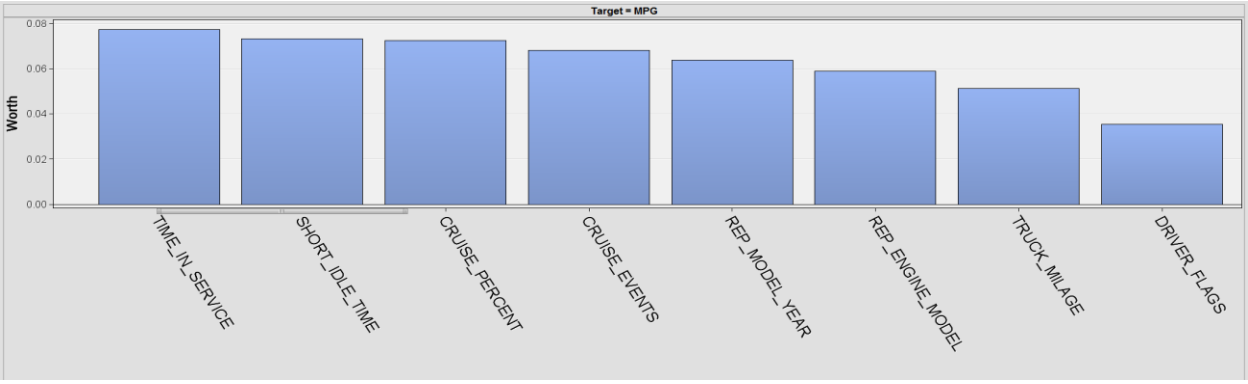
Linehaul Nominal Variables

Only 4 nominal variables contain enough variation to necessitate further investigation to rule out confounding variables. The REP_ENGINE_MAKE_TXT, REP_ENGINE_MODEL, REP_MODEL, REP_MODEL_YEAR variables will all be further investigated in their relation to the target variable. 12 of the 16 nominal variables are > 90% uniform and thus will not introduce confounding factors into the analysis to come.

Variable	Worth
TIME_IN_SERVICE	0.077352
SHORT_IDLE_TIME	0.073106
CRUISE_PERCENT	0.072482
CRUISE_EVENTS	0.068133
REP_MODEL_YEAR	0.063741
REP_ENGINE_MODEL	0.059016
TRUCK_MILAGE	0.051127
DRIVER_FLAGS	0.035452

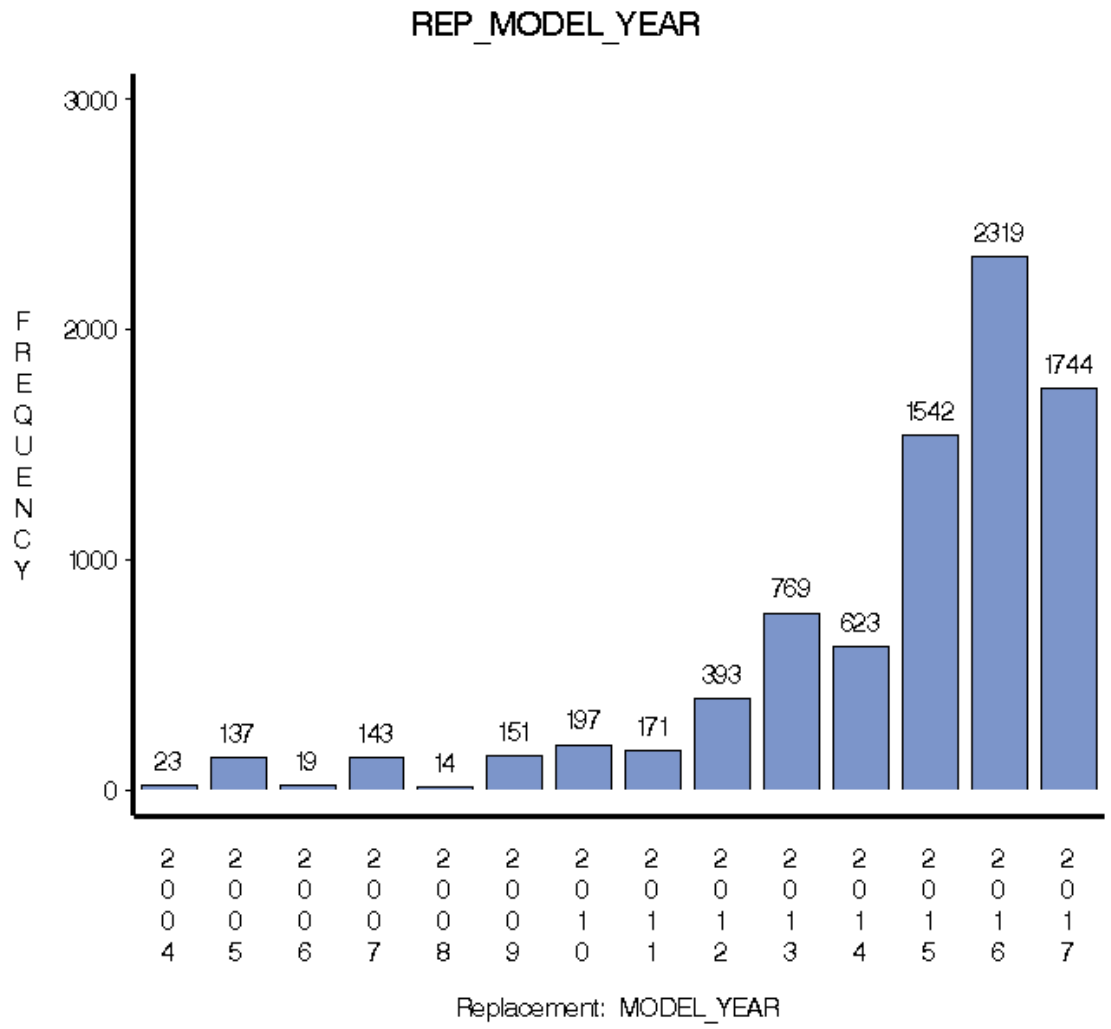
3 Significant Linehaul Variable Worth

Using the variable worth calculated by the Stat Explore node, it can be appreciated that of the 4 nominal variables with the first mode occurring in < 90% of observations, only REP_MODEL_YEAR and REP_ENGINE_MODEL have a worth greater than .05. Thus the relation between these two variables and the target MPG variable will be further investigated



4 Linehaul Variable Worth

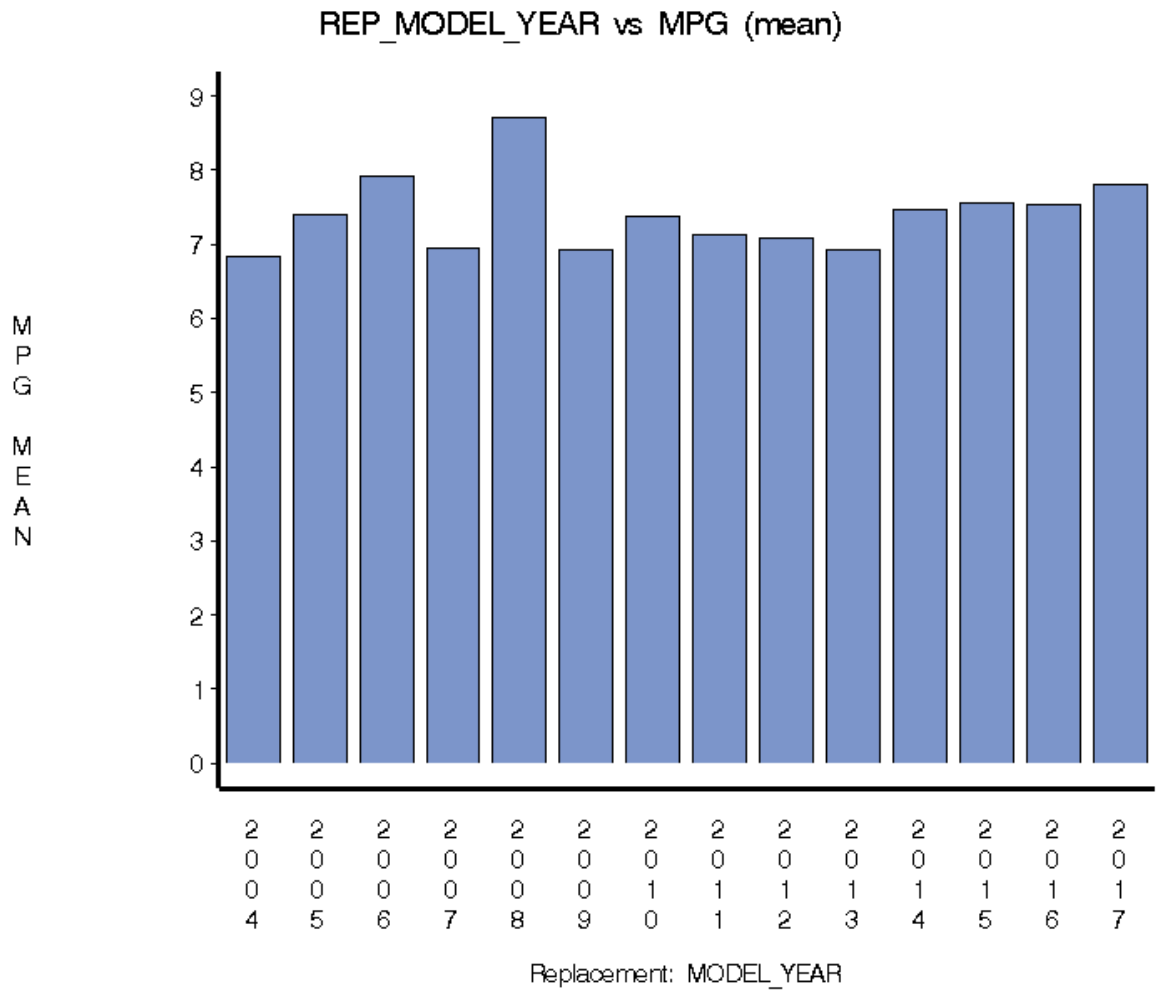
. Model Year



5 Linehaul Frequency by Model Year

With a worth of .06, the REP_MODEL_YEAR variable has 14 levels. With 2016 being the most common and 2017 the second with tractors becoming less common as they get older for the most part after that.

This distribution is heavily left skewed though that is expected as the company ages out tractors as they become older. Note that all trucks of 2011 model year and before have less than 200 instances within the data set.

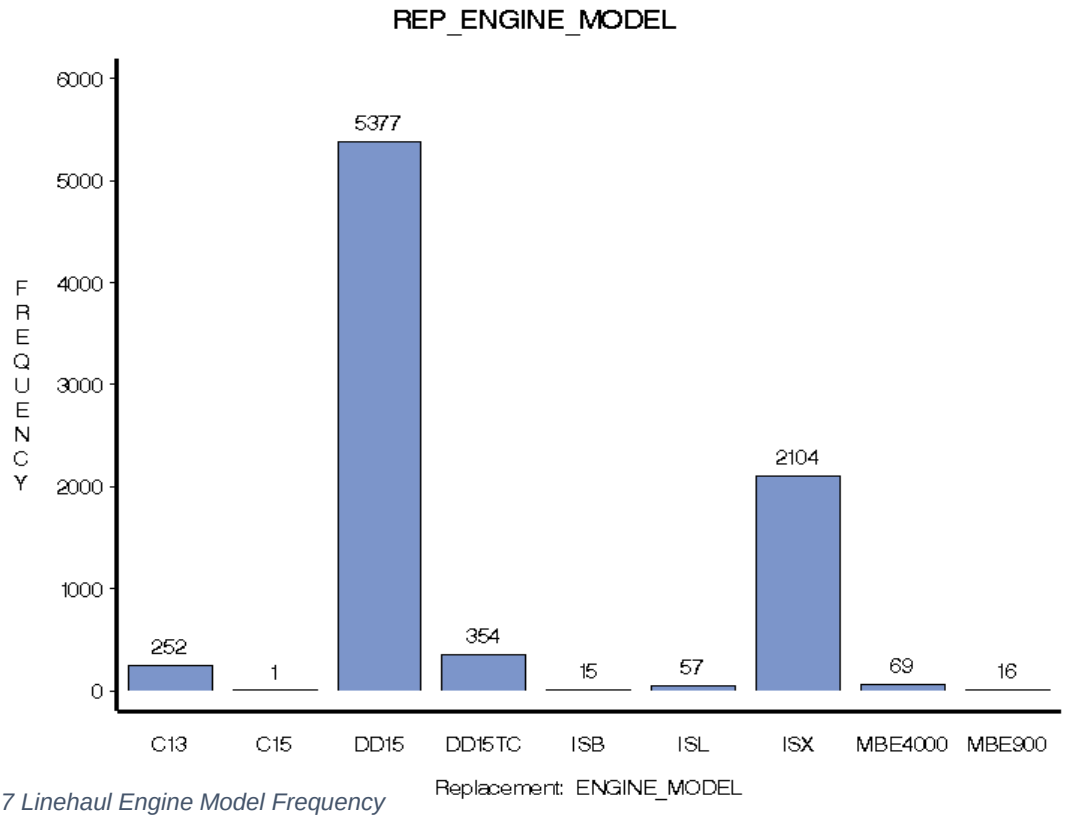


6 Linehaul Average MPG by Model Year

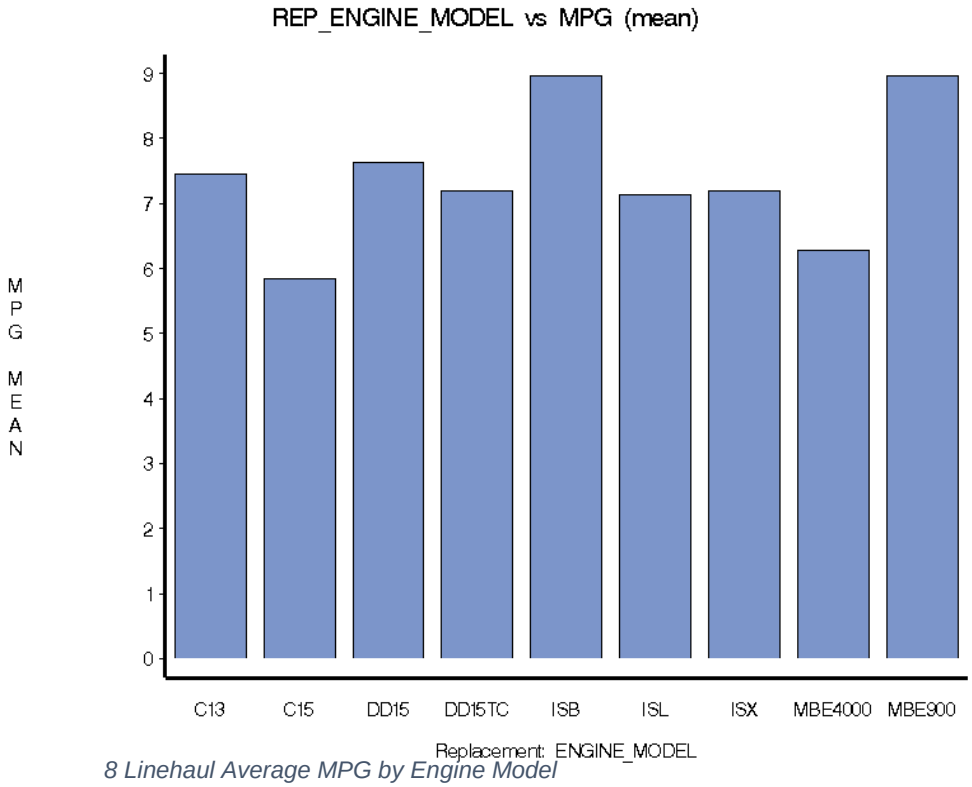
Comparing MODEL_YEAR to MPG, it can be appreciated that most trucks return about 7 MPG, though models 2014 and newer seem to have an improved fuel efficiency with spikes during the 2008 and 2006 model years. These spikes can easily be attributed to a small sample size as the 2006 and 2008 model years combined only make up 33 observations.

An investigation as to why models 2013 and newer would be an interesting study though that is outside of the scope of this report. For this investigation, the model year of the truck being driven will need to be included within the assumptions and margin for error of the final product though drivers are rarely able to select which trucks they are to operate.

Engine Model



7 Linehaul Engine Model Frequency



8 Linehaul Average MPG by Engine Model

In these graphs, a couple conclusions can be drawn. First, within the dataset, the DD15 and ISX engine models are make up 92% of all engines. This means that data from each of the other engine types can easily be skewed due to sample size and thus must be disregarded. We see that the difference of .3 MPG between the DD15 engine and the ISX engine could be statistically significant and thus must be kept in mind when evaluating model error.

Linehaul interval variables

Figure 9 contains information detailing the linehaul interval variables.

Variable	Mean	Standard Deviation	Non Missing	Min	Median	Max	Skewness	Kurtosis
BRAKE_EVENTS	60.55	64.21003	8245	-1	36	599	1.988352	5.551886
CRUISE_EVENTS	4.052	7.382251	8245	0	1	142	4.990795	48.89899
CRUISE_PERCENT	0.292	0.328886	8245	0	0.131443	1.002801	0.646523	-1.11784
DRIVER_FLAGS	103.2	46.69475	8245	6	102	253	0.37751	-0.50297
EXCESS_SPEED	16.68	69.16191	8245	0	0	1332	8.148081	88.72943
JAKE_BRAKE	76.97	225.0372	8245	0	1	2459	5.395902	35.12352
LONG_IDLE_TIME	141.7	627.0245	8245	0	0	11214	6.842662	64.47012
OVER_RPM	52.1	125.0225	8245	0	6	1927	4.80244	33.19297
OVER_SPEED	39.44	152.1579	8245	0	0	2210	8.222465	85.8265
SEATBELT_TIME	540.2	2337.252	8245	0	0	20761	4.959522	25.7482
SHORT_IDLE_TIME	554.6	786.8217	8245	0	250	7398	2.627465	9.288169
TIME_IN_SERVICE	1357	925.4235	7749	457	1154	4906	1.725831	3.195962
TIME_SINCE_HIRE	7.946	5.039612	8245	1.988	6.525	29.931	0.89667	0.218502
TIME_TO_FULL_TIME	0.382	1.678819	8238	0	0	20.647	7.228084	62.68069
TOPGEAR_PERCENTAGE	0.01	0.089905	8245	0	0	0.995206	9.215407	85.98194
TRUCK_MILAGE	4E+05	239511.4	8245	0.2	400892.2	1073311	0.396629	-0.491
MPG	7.481	1.241536	8245	2.078	7.526	18.254	-0.09432	3.409391

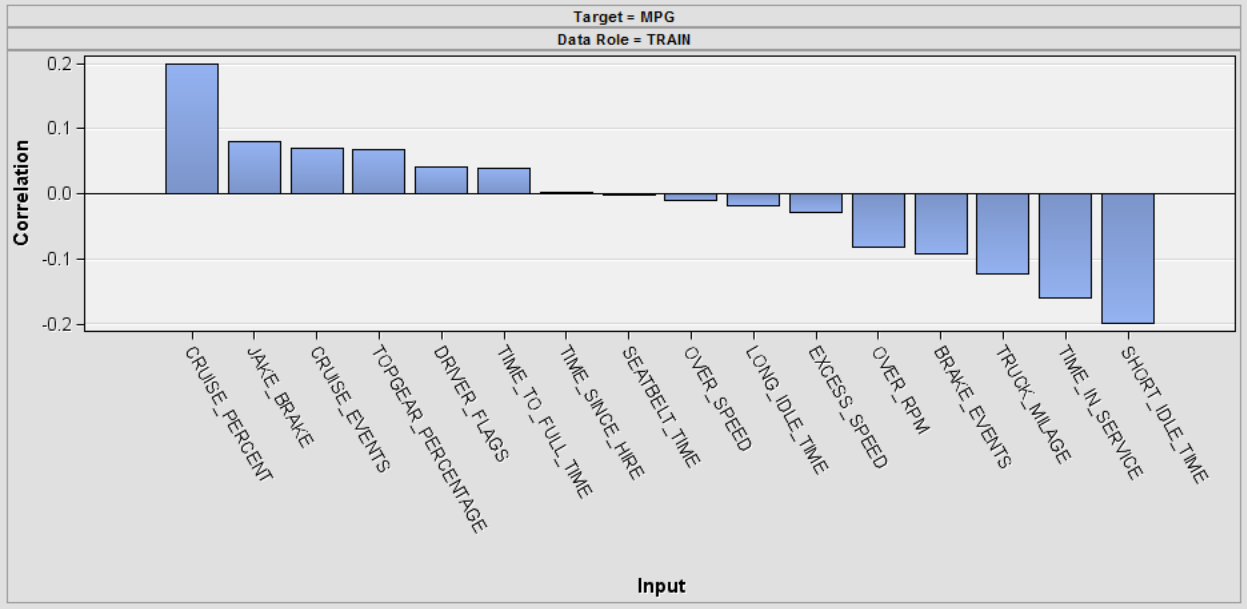
9 Linehaul Interval Variables

Multiple points of interest are available within this output.

1. The median and mode of the target are within .04 standard deviations of one another indicating a near center
2. A skewness of near zero and kurtosis of 3.4 indicate that this set's target variable is symmetric with light tails. This indicates a near normal distribution when taken in context with point 1.
3. A total of 503 data points are missing. 496 TIME_IN_SERVICE and 7 TIME_TO_FULL_TIME points. These missing values are handled in different ways depending on the model. This is discussed in a later section.
4. Mean Cruise percentage is .29 though a standard deviation of .33 indicates this variable is not normally distributed.
5. All dispatches contained at least 6 driver flags with an average of 103.2 and maximum of 253 flags

6. All variables except MPG are skewed right to varying degrees
7. Average truck mileage is 407909.6 with a minimum of .2 and the longest tenured truck having 1073311 miles
8. The average driver has been employed by Old Dominion Freight Line for 7.9 years with the skew being slightly to the right though the median falls just short of the mean at 6.5 years.
9. SHORT_IDLE_TIME seems to be skewed by outliers with the mean (554.5) over double the median of 250.
10. OVER_RPM median is 6 while the mean is 52
11. OVER_SPEED median is 0 while the mean is 39.4
12. TOPGEAR_PERCENTAGE median is 0 while the mean is .01 with a relatively large standard deviation of .09

To determine which variables have the most effect on the target variable, we take a look at the Pearson Correlation Plot and the ensuing table of values.



10 Linehaul Pearson Correlation

Input	Correlation
CRUISE_PERCENT	0.19939
JAKE_BRAKE	0.07969
CRUISE_EVENTS	0.07072
TOPGEAR_PERCENTAGE	0.06769
DRIVER_FLAGS	0.04046
TIME_TO_FULL_TIME	0.03895
TIME_SINCE_HIRE	0.00158
SEATBELT_TIME	-0.00108
OVER_SPEED	-0.01031
LONG_IDLE_TIME	-0.01763
EXCESS_SPEED	-0.02875
OVER_RPM	-0.08235
BRAKE_EVENTS	-0.09306
TRUCK_MILAGE	-0.12276
TIME_IN_SERVICE	-0.1607
SHORT_IDLE_TIME	-0.19931

11 Linehaul Correlation Coefficients

Figures 10 and 11 indicate that the strongest positive correlations are resulting from the CRUISE_PERCENT, JAKE_BRAKE, CRUISE_EVENTS, and TOPGEAR_PERCENTAGE variables while the strongest negative correlations are resulting from the SHORT_IDLE_TIME, TIME_IN_SERVICE, TRUCK_MILAGE, BRAKE_EVENTS, and OVER_RPM variables.

In short, this indicates that if the driver can utilize the cruise control, the top-most gear of the truck, and the Jake brake while minimizing the amount of short idling and brake/over rpm events then their MPG should be maximized. However this does also indicate that the longer the truck is in service the worse the MPG should be expected to be so this must be taken into consideration.

PND descriptive analysis

A second, separate descriptive analysis detailing the PND data set is conducted next.

PND Nominal Variables

In the table following, the output from the Stat Explore node for the nominal variables is shown

Variable Name	Levels	Missing	Mode	Mode Percentage	Mode2	Mode 2 Percentage
REP_CAB_TYPE	3	0	DAYCAB	90.55	SLEEPER	7.69
REP_DRIVE_AXLE_SET_UP	2	0	SINGLE AXLE	96.49	TANDEM AXLE	3.51
REP_ELOG_CERTIFIED_FLG	3	0	Y	97.62	N	2.22
REP_ELOG_TRAINED_FLG	3	0	Y	97.62	N	2.22
REP_ENGINE_MAKE_TXT	6	0	DETROIT	58.82	CUMMINS	29
REP_ENGINE_MODEL	11	0	DD15	55.26	ISX	25.32
REP_EQUIPMENT_CATEGORY_TYPE_NM	4	0	DAYCAB	89.41	SLEEPER	7.69
REP_FIFTH_WHEEL_TYPE	2	0	FIXED	95.66	SLIDE	4.34
REP_FULL_TM_FLG	2	0	F	99.79	P	0.21
REP_LINEHAUL_PACOS_TRAINED	3	0	Y	97.62	N	2.22
REP_MAKE	3	0	FRGHT	90.17	VLVNA	9.8
REP_MODEL	11	0	CA125	74.5	CL112	10.78
REP_MODEL_YEAR	14	0	2016	27.24	2015	17.28
REP_PND_PACOS_TRAINED	2	0	N	99.84	U	0.16
REP_REAR_AXLE_RATIO	10	0	3.42	89.1	3.58	5.46
REP_TRANSMISSION_MAKE	5	0	EATON	85.71	MERTR	9.12

12 PND Nominal Variable Summary Statistics

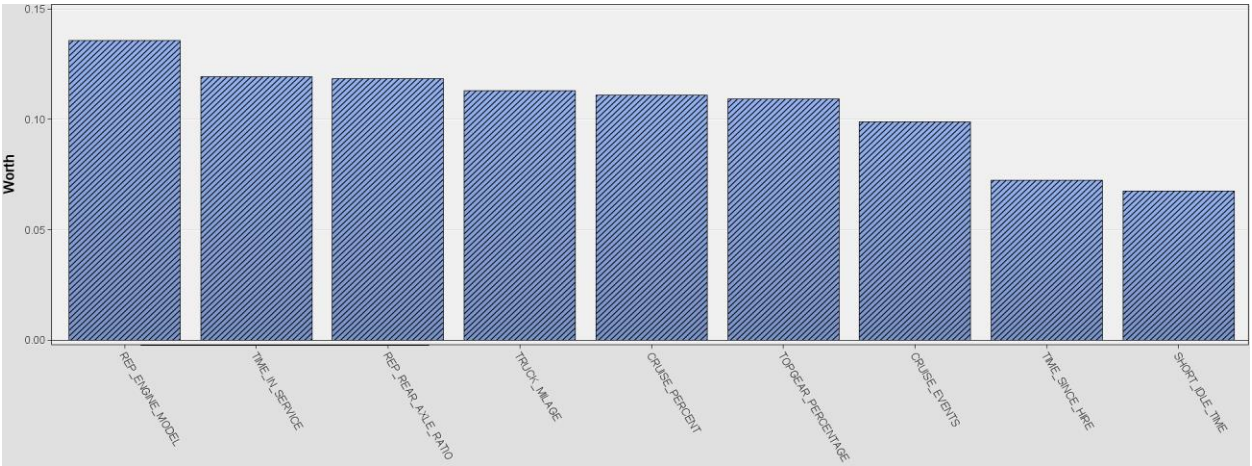
Following this output, these observations can be made

1. 7 variables have a mode that appears in less than 90% of the data points (REP_ENGINE_MAKE_TXT, REP_ENGINE_MODEL, REP_EQUIPMENT_CATEGORY_TYPE_NM, REP_MODEL, REP_MODEL_YEAR, REP_REAR_AXLE_RATIO, REP_TRANSMISSION_MAKE)
2. Only about 90% of PND routes use day cabs
3. Over 99% of drivers are full time
4. 99.8% of PND drivers have not received the PND PACOS training while 97.6% have received the linehaul PACOS training

5. 90% of tractors in this set are made by Freightliner
6. Like with the linehaul, the most common year model is 2016 but unlike the linehaul set, the next most common year model is 2015
7. The most common engine manufacturer is Detroit at 58.8% of all tractors in the set

With these points in mind, a further look at the 7 nominal variables with higher degrees of variability is required.

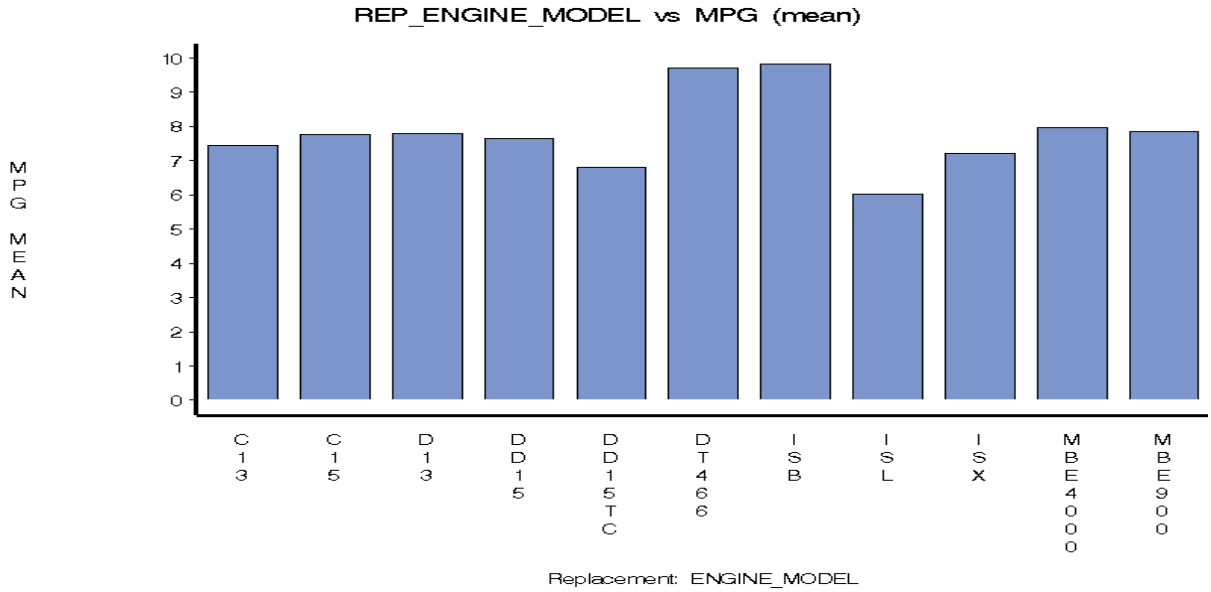
In looking at the variable worth, only two of the seven nominal variables with modes appearing in less than 90% of the data are identified as having a worth greater than .05. REP_ENGINE_MODEL has a worth of .136 and REP_REAR_AXLE_RATIO has a worth of .118



13 PND Variable Worth

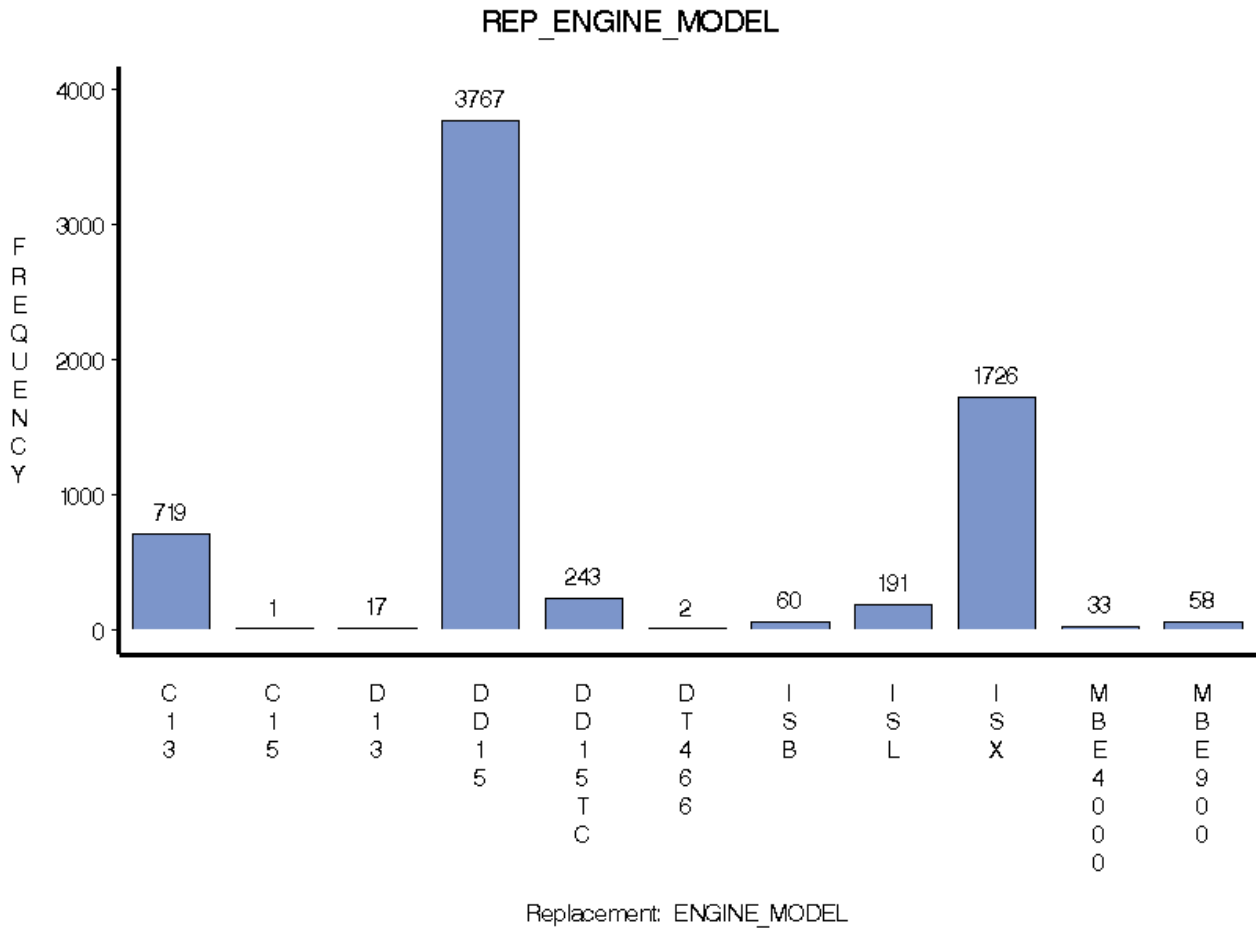
Engine Model

Graphing the REP_ENGINE_MODEL variable by the target MPG variable the following graph is obtained.

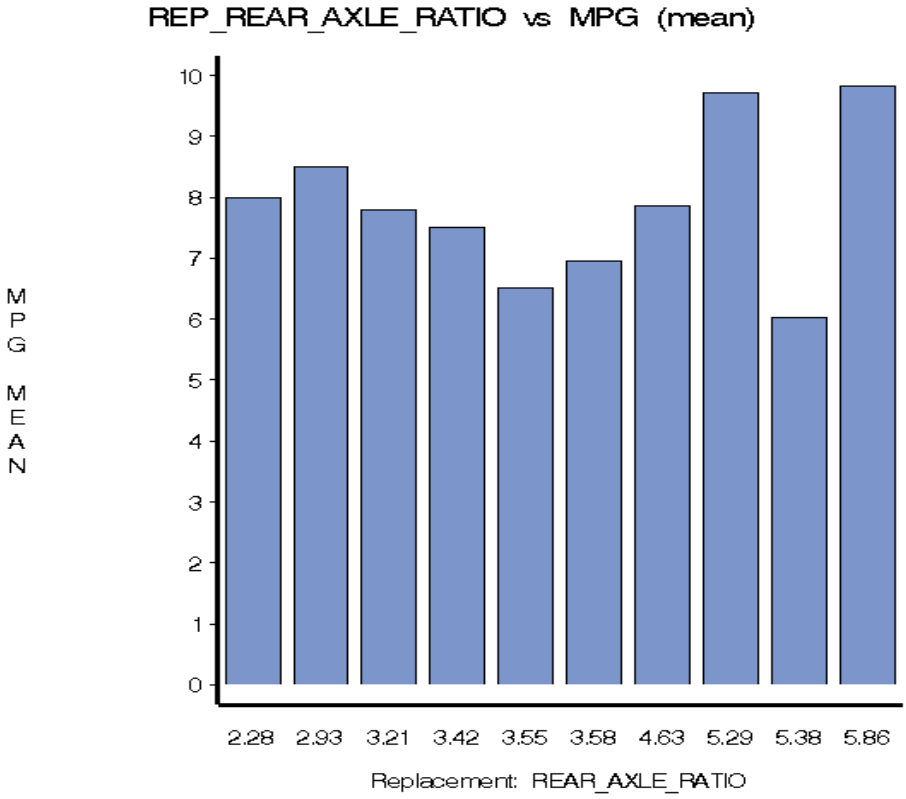


14 PND Average MPG by Engine Model

Most engine models return between 7.5 and 8.1 MPG with the DD15TC and ISL returning an average MPG in the 6 to 6.9 MPG range. Conversely the DT466 and ISB engine models seem to perform abnormally well with both returning a mean MPG of 9+ MPG.



15 PND Engine Model Frequency

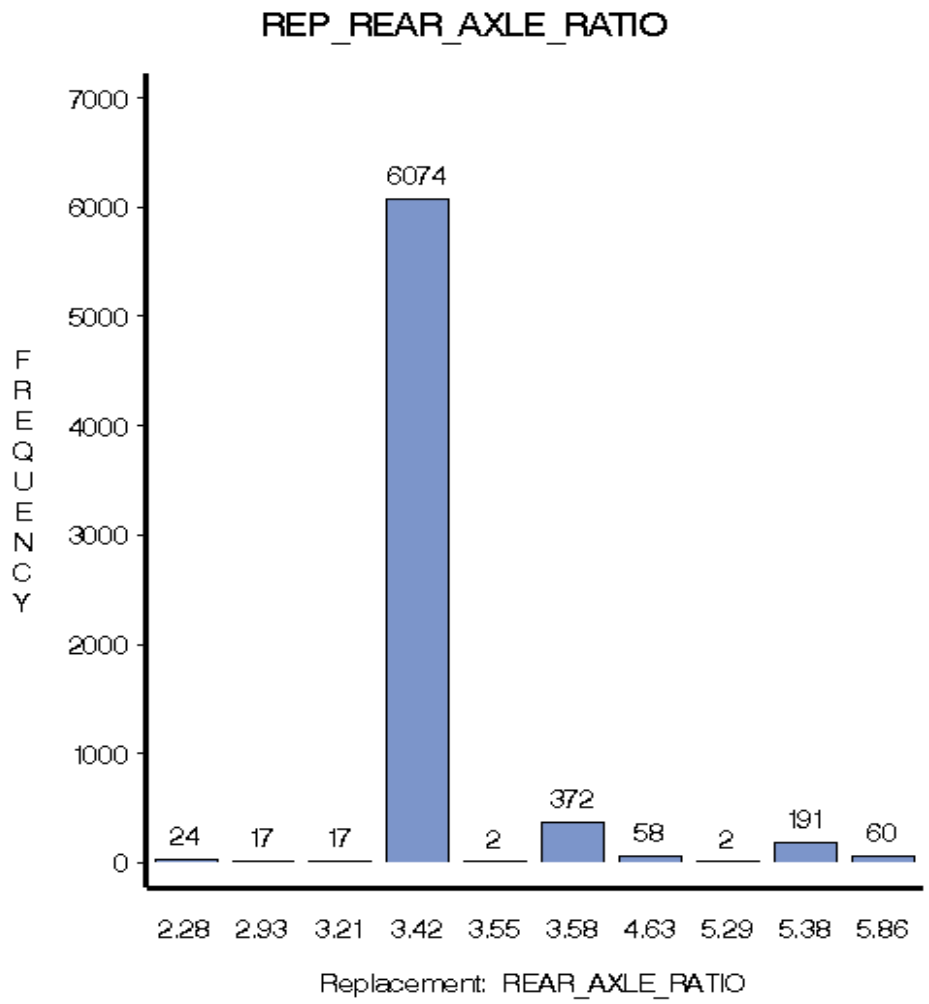


16 PND MPG by Rear Axle Ratio

Next by looking at the frequency graph of the REP_ENGINE_MODEL variable it becomes clear that due to sample size, only the C13, DD15 and ISX models return a reliable MPG estimate. All other engine models fail to reach even 250 data points. It should be noted however, that of the three reliable levels, the DD15 returned the highest average PG if even by only a few tenths of a mile.

Rear Axle Ratio

When comparing the mean MPG returned by each rear axle ratio, we see that most rear axle ratios return an average MPG from ± 1 standard deviation ($6.2 < x < 8.7$) from the sample mean with all outliers falling to the rear axle ratios greater than 4.63.



17 PND Rear Axle Ratio Frequency

Next, quickly looking at the frequency graph of the REP_REAR_AXLE_RATIO shows that while the 3.42 rear axle ratio doesn't quite meet the 90% threshold, no other variable level surpasses 372 data points and thus most if not all variability can be attributed to small sample sizes.

PND Interval Variables

Now we examine the interval variables within the PND partition. As is the nature of PND routes, more quick-stop opportunities are expected while the usefulness of the cruise control and top gear will be minimized. See the output of the summary statistics below

Variable	Mean	Standard Deviation	Non Missing	Minimum	Median	Maximum	Skewness	Kurtosis
BRAKE_EVENTS	89.1983	81.76738	6817	-1	67	715	1.601356	3.60788
CRUISE_EVENTS	2.64969	9.410465	6817	0	0	682	55.58876	3986.37
CRUISE_PERCENT	0.19227	0.270497	6817	0	0	1.047619	1.256356	0.34697
DRIVER_FLAGS	80.3315	46.79537	6817	1	59	505	1.694717	6.40648
EXCESS_SPEED	9.82939	44.26688	6817	0	0	860	8.033187	88.3057
JAKE_BRAKE	42.8716	158.0929	6817	0	0	2983	7.447715	74.8925
LONG_IDLE_TIME	130.297	550.7645	6817	0	0	8393	6.330642	50.2137
OVER_RPM	40.9906	111.7479	6817	0	5	1969	6.635777	65.0137
OVER_SPEED	20.8390	131.8739	6817	0	0	9159	49.82708	3387.50
SEATBELT_TIME	739.222	2825.588	6817	0	0	26389	4.475223	21.2142
SHORT_IDLE_TIME	864.203	911.1901	6817	0	562	7738	1.623182	3.49290
TIME_IN_SERVICE	1746.56	1237.283	6326	302	1266	5021	1.15894	0.31624
TIME_SINCE_HIRE	9.11904	6.412173	6817	1.961	7.04	41.717	1.336481	2.63285
TIME_TO_FULL_TIME	0.26151	0.986988	6803	0	0	10.874	7.212222	62.4079
TOPGEAR_PERCENTAGE	0.02156	0.118389	6817	0	0	1	5.980927	36.5485
TRUCK_MILAGE	444174	267945.1	6817	11.5	450612.5	1113806	0.117885	-0.79232
MPG	7.46520	1.444042	6817	1.775	7.508	19.841	0.116937	2.69446

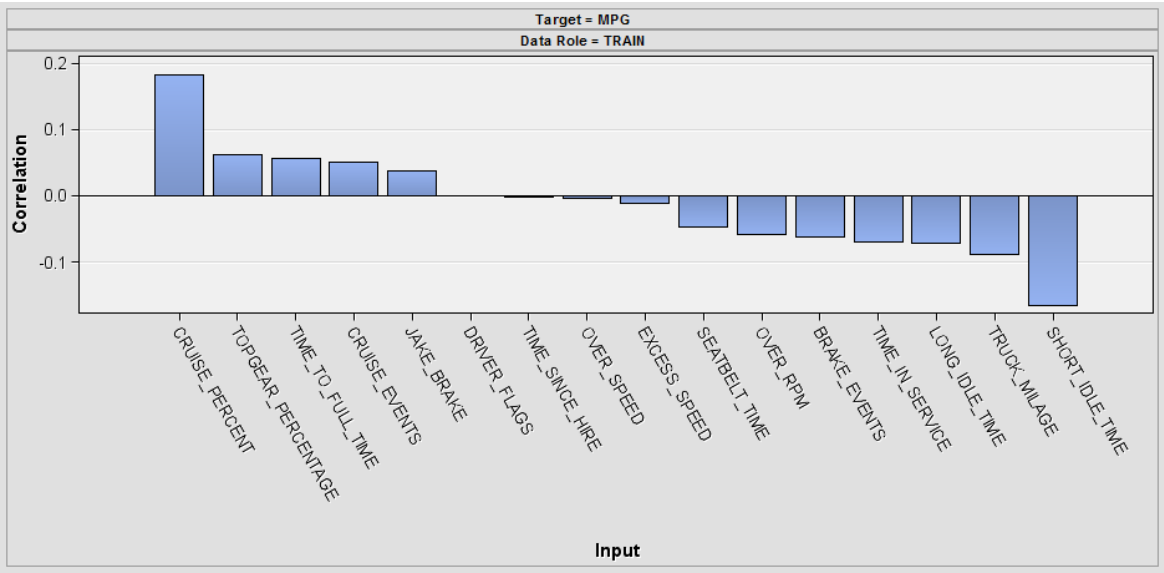
18 PND Interval Variable Summary Statistics

1. MPG mean is 7.5 with a standard deviation of 1.44. Combined with a skewness of .117 and a kurtosis of 2.69, we can say the MPG variable is relatively normal with a slight skew to the right and slightly steeper decline as we move away from the center than a perfect normal curve.
2. Mean brake events are up to 89.2 with a standard deviation of 81.8.

3. Mean cruise percent is .19 with a standard deviation of .27 while top gear percentage is also down to a .02 mean
4. Mean driver flags is 80 with a minimum of 1 and max of 505. A median of 55 indicates the curve is heavily skewed right
5. Mean SHORT_IDLE_TIME is up to 864.2 while the mean LONG_IDLE_TIME is down to 130.3
6. TIME_SINCE_HIRE is up to 9.11 from 7.94 in the linehaul set indicating more experienced drivers
7. Most variables are still heavily skewed to the right with heavy tails as indicated by a positive skewness value and kurtosis values greater than three.
8. TIME_IN_SERVICE increases from 1357 in the linehaul dataset to 1746 in the PND dataset

Overall, the dispatches in the PND dataset return roughly the same MPG average but do so with more experienced drivers in older trucks who have less opportunity to use their cruise control or top gear and are subjected to heavy braking events more often. This is interesting and indicates that the more experienced drivers should have a heavy effect on the MPG returned.

When viewing the Pearson Correlation plot however, we still see that Cruise events have a heavily positive impact on the target variable along with TOPGEAR_PERCENTAGE, TIME_TO_FULL_TIME, CRUISE_EVENTS, and JAKE_BRAKE. While the strongest negative correlations belong to SHORT_IDLE_TIME, TRUCK_MILAGE, LONG_IDLE_TIME, and TIME_IN_SERVICE. This correlation plot is much heavier skewed to the negative side indicating that there are more opportunities to bring the MPG down than there are to bring it up. This is expected with these shorter more intermittent routes being ran by the PND drivers.



19 PND Interval Variable Pearson Coefficient

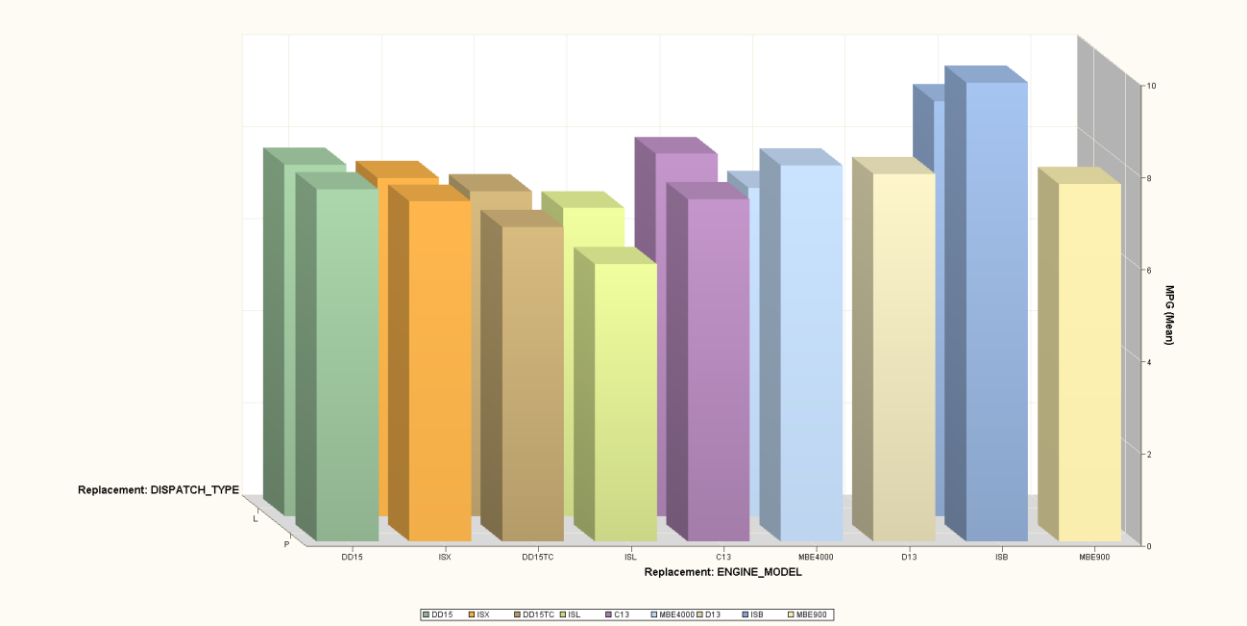
Input	Correlation
CRUISE_PERCENT	0.1818
TOPGEAR_PERCENTAGE	0.0618
TIME_TO_FULL_TIME	0.05669
CRUISEEVENTS	0.05075
JAKE_BRAKE	0.038399
DRIVER_FLAGS	0.0002
TIME_SINCE_HIRE	-0.00175
OVER_SPEED	-0.00418
EXCESS_SPEED	-0.01088
SEATBELT_TIME	-0.04632
OVER_RPM	-0.0582
BRAKE_EVENTS	-0.06123
TIME_IN_SERVICE	-0.07006
LONG_IDLE_TIME	-0.07052
TRUCK_MILAGE	-0.08882
SHORT_IDLE_TIME	-0.16599

20 PND Pearson Correlation Coefficient

Of interesting note however, the TIME_SINCE_HIRE variable has a slightly negative effect on the MPG returned. This slightly negative correlation indicates that as the time since being hired increases, the average MPG returned for a driver's routes tends to decrease. Possible explanations for this will be provided later. The correlation coefficients can be seen in figure 20.

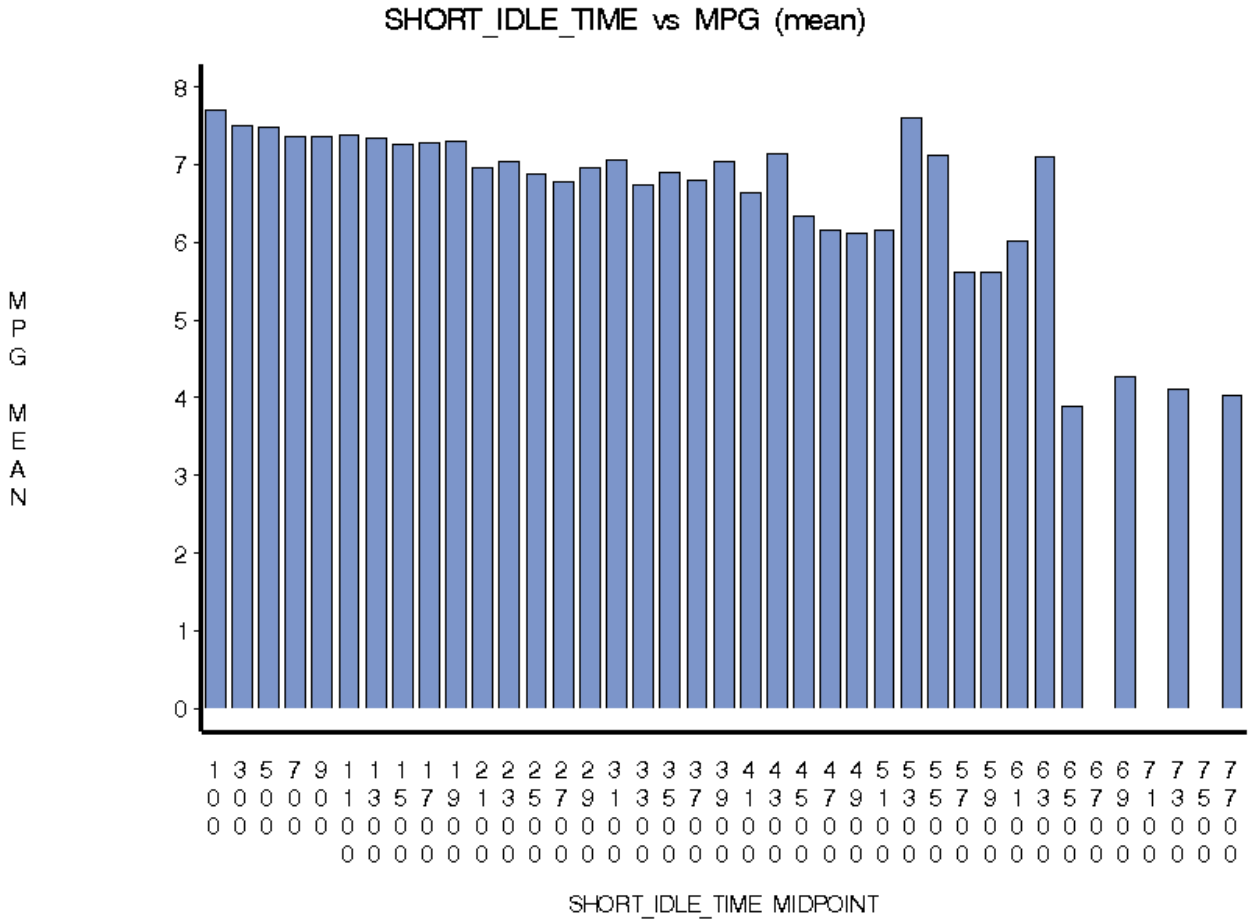
Important variables across the population

From the descriptive analysis of both the PND and linehaul partitions, it was identified that the ENGINE_MODEL variable has worth in all routes run by the company. In performing a comparison of each ENGINE_TYPE response we can appreciate in figure 21 that the ISL and C13 engine types tend to perform best in a linehaul route while the MBE4000 and the ISB tend to perform best as a PND engine.



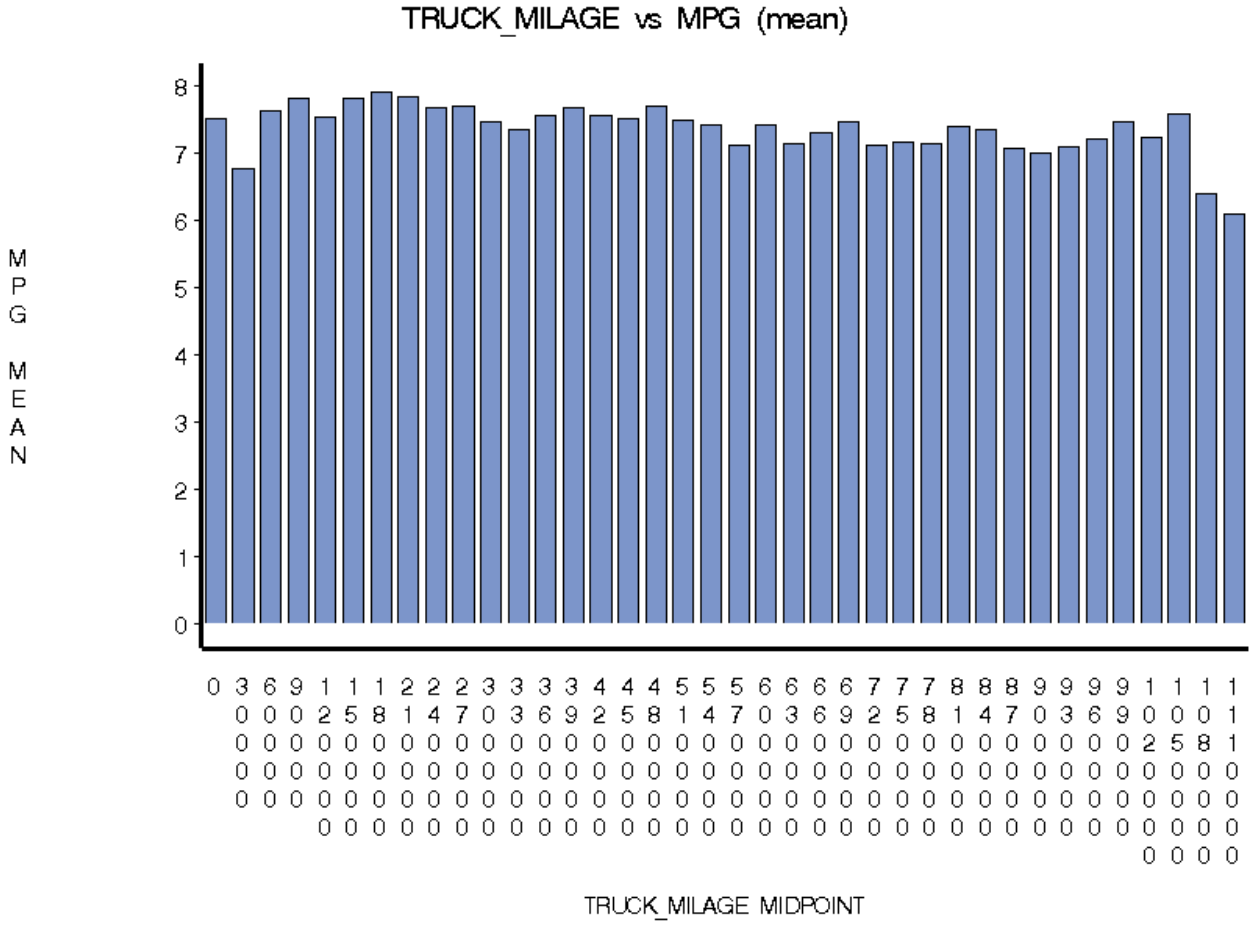
21 Population MPG by Engine Model and Dispatch Type

Additionally the SHORT_IDLE_TIME variable was indicated to have a negative correlation in both the linehaul and PND sets. Figure 22 shows that this correlation holds true when addressing the entire set. This is to be expected as short idle times are defined as when a tractor is left to idle for less than 300 seconds or 5 minutes. These generally indicate waiting at a stoplight/sign or some other form of traffic. Linehaul routes are not immune to short idle times as demonstrated previously however short idle times are more expected within the PND dataset. Below is a chart demonstrating the effect that extended short idle times has on the MPG. Notice the slight but generally negative trend.



22 Population Average MPG by Short Idle Time

The last variable that may cause an issue in both the PND and linehaul datasets is the TRUCK_MILAGE variable. The hypothesis being that older trucks should have lower gas mileage than newer trucks. Whether this is a function of the wear and tear sustained by the truck or of the advancements made on new trucks to improve MPG is outside the scope of this project. We can acknowledge that this hypothesis tends true within the dataset. As the graph below moves to the right, the MPG slowly decays though not to a generally significant amount.



23 Population MPG by Truck Mileage Midpoint Bucketing

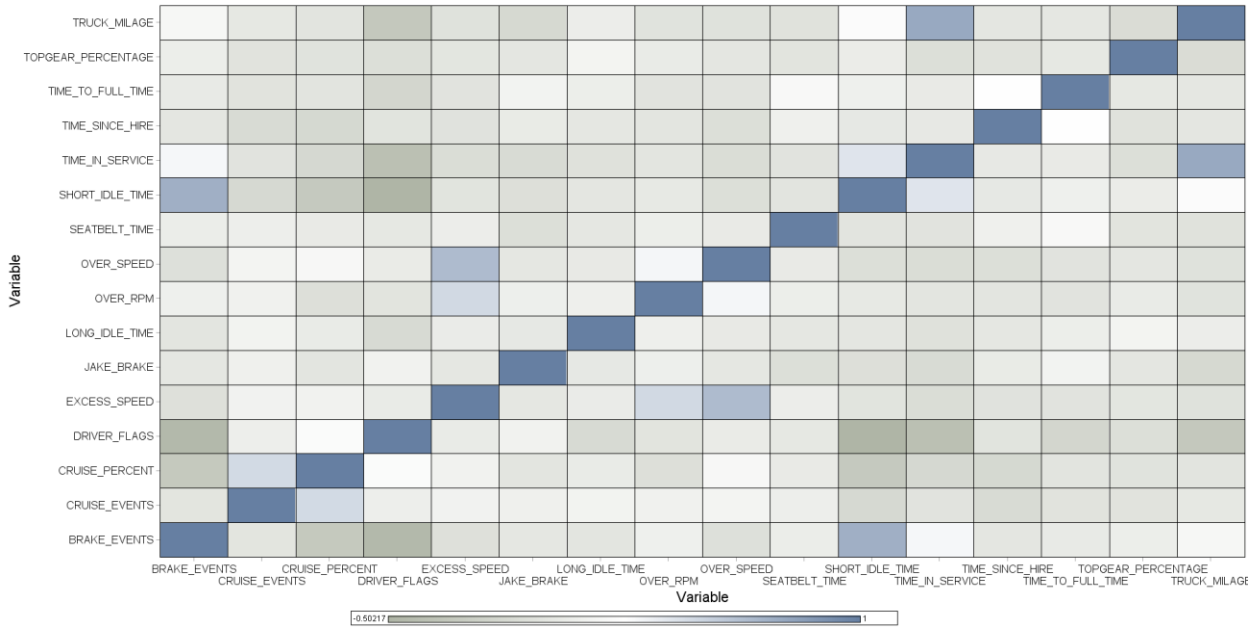
Correlation analysis

Looking at the dataset we try to determine how closely related each variable is to the target variable of MPG. Why do they move in the same or opposite directions and which variables should be expected to not affect the target at all?

Determining the predictor variables for a given target is the major challenge when performing predictive analytics and as such we try to use data descriptive of our target. In this case, this is data considering: driver history, driver inputs during dispatch, route information, descriptive information about the truck being used, and closely related data points.

In order to answer these questions, a correlation analysis and descriptive scatter plots between observed MPG and all collected variables during the June 2017 timeframe are produced.

To see how all variables relate to one another, we produce the correlation matrix in figure 24.



24 Correlation Matrix

Variable	Variable2	Correlation
TRUCK_MILAGE	TIME_IN_SERVICE	0.752200593
SHORT_IDLE_TIME	BRAKE_EVENTS	0.716423017
OVER_SPEED	EXCESS_SPEED	0.645020062
OVER_RPM	EXCESS_SPEED	0.472952347
CRUISE_PERCENT	CRUISE_EVENTS	0.468147473
TIME_IN_SERVICE	SHORT_IDLE_TIME	0.40569834
DRIVER_FLAGS	BRAKE_EVENTS	-0.463419039
DRIVER_FLAGS	SHORT_IDLE_TIME	-0.50216757

25 Significant Correlation Pairs

Though most of the input variables fail to register as either strongly positive or negatively correlated, there are a few pairs that have some correlation. In figure 25 is the listing of all variable pairs with Pearson Correlation value greater than .4 or less than -.4

As expected, a longer time in service indicates a given truck has higher mileage, the more brake events that are registered the more time a truck will spend idling, and the more times the excess speed threshold is crossed the more often the over speed and over rpm threshold will be crossed. The final three rows hold the most interesting information. The longer a truck is in service the more likely it is to have short idle times. This is evident from our discovery that that the average age of PND trucks is much greater than the average age for linehaul trucks. Since PND trucks are more likely to be stopped for brief periods of time during their routes this correlation makes sense. Also the more driver flags that are registered, the less brake events and short idle times are registered. This may be a clustering to perform more research on in future projects.

In addition to this input variable correlation analysis, we will use a variable selection node to reduce the number of inputs when beginning the predictive analytics phase for certain models. This node uses Chi-square and R-square selection criterion to remove variables that pose little to no effect on the target variable. This variable selection will be performed on both the linehaul and PND datasets.

The results of the Variable Selection node for the linehaul dataset are seen below.

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Group:REP_MODEL_YEAR	4	0.05004	108.5128	<.0001	635.8804	1.464989
Var:CRUISE_PERCENT	1	0.030934	277.322	<.0001	393.0917	1.417456
Var:SHORT_IDLE_TIME	1	0.012831	116.6393	<.0001	163.0427	1.397836
Group:REP_ENGINE_MODEL	3	0.013946	42.90444	<.0001	177.2158	1.376826
Var:BRAKE_EVENTS	1	0.00709	65.94904	<.0001	90.0898	1.366052
Class:REP_ENGINE_MAKE_TXT	3	0.007367	23.02562	<.0001	93.61133	1.355177
Var:OVER_RPM	1	0.006518	61.56863	<.0001	82.82688	1.345277
Class:REP_DRIVE_AXLE_SET_UP	1	0.003868	36.69181	<.0001	49.14751	1.339468
Group:REP_REAR_AXLE_RATIO	3	0.003262	10.35185	<.0001	41.45659	1.334917
Var:JAKE_BRAKE	1	0.001864	17.77811	<.0001	23.68399	1.3322
Class:REP_EQUIPMENT_CATEGORY_TYPE_NM	2	0.001176	5.617349	0.0036	14.95008	1.330706
Var:CRUISE_EVENTS	1	0.001053	10.07122	0.0015	13.38706	1.329239
Group:REP_MODEL	4	0.000693	1.656476	0.1571	8.804602	1.328815

26 Selected Linehaul Variables

The resulting 13 variables are all considered pertinent to the MPG target due to having either a p-value or R-Square of less than .05. Of the 13, the first 12 variables have a p-value of less than .05 and the REP_MODEL group is retained because of its exceedingly low R-Square value. These variables are what we will use during the predictive modelling phase.

Following are the results for the variable selection node for the PND partition.

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Group:REP_ENGINE_MODEL	5	0.078127	115.443762	<.0001	1110.42626	1.923753
VAR:CRUISE_PERCENT	1	0.033122	253.79545	<.0001	470.76682	1.854906
Group:REP_MODEL_YEAR	5	0.023582	37.097335	<.0001	335.177167	1.807015
Var:LONG_IDLE_TIME	1	0.010247	81.551477	<.0001	145.640762	1.785875
Var:SHORT_IDLE_TIME	1	0.007697	61.808086	<.0001	109.403776	1.770056
Var:BRAKE_EVENTS	1	0.007127	57.702234	<.0001	101.291932	1.755425
Group:REP_MODEL	5	0.005679	9.252456	<.0001	80.720285	1.74484
Var:TOPGEAR_PERCENTAGE	1	0.005388	44.170558	<.0001	76.584149	1.733828
Group:REP_TRANSMISSION_MAKE	3	0.01574	43.823151	<.0001	223.716353	1.70166
Var:OVER_RPM	1	0.00395	33.150537	<.0001	56.145224	1.693644
Var:TIME_TO_FULL_TIME	1	0.002268	19.08662	<.0001	32.240095	1.689146
Var:CRUISE_EVENTS	1	0.000817	6.877605	0.0087	11.607236	1.687686
Var:SEATBELT_TIME	1	0.000751	6.327973	0.0119	10.671257	1.686362

27 Selected PND Variables

The requirements for variable selection had to be altered for the PND dataset. When the standard parameters for variable selection were left as default, only the REP_ENGINE_MODEL, CRUISE_PERCENT, REP_MODEL_YEAR, and LONG_IDLE_TIME variables were selected. We did not feel these four variables were representative of the dataset and thus the selection criteria for the minimum R-Square was lowered from .005 to .002. This resulted in the 13 variables above being selected.

Preparing the data for modelling

For this set of data, we compared several permutations spanning four different types of models. The model types compared were Regression, Decision Tree, Neural Network and Memory Based Reasoning (Nearest Neighbor) models. For each group, the data must be prepared slightly differently because of how they handle data.

Before beginning modeling, both the PND and linehaul datasets must be partitioned into a training and validation set.

The training set will be used to develop the models after which the models will be applied to the validation set for model. Using the data partition node, we will create a 65-35 split between training and validation sets. This results in the following datasets.

	Number of Observations	
	Linehaul	PND
Training	5359	4431
Validation	2886	2386

Decision Tree Preparation

The Decision Tree branch of the model comparison is simple. Decision trees are able to handle missing values by using surrogate rules in place of the missing values to estimate what those values would be based on similar observations. Decision trees are also relatively unaffected by extreme values so normalization is unnecessary. Decision trees are made by using splitting rules to separate the data repeatedly into similar groups using variables with high value to the target variable. The trees created for this project require a minimum leaf size of 5 and a maximum depth of 6 iterations to limit complexity.

Neural Network Preparation

Neural Networks are machine learning algorithms that are typically most useful when dealing with very large datasets containing many non-linear relationships. A major drawback of Neural Networks however are that they are susceptible to over-training. In an attempt to alleviate this issue, the variable selection results from earlier will be used to inform the neural networks. This will reduce the number of input variables for the neural network to use and subsequently reduce the likelihood of overtraining on the training dataset.

Memory Based Reasoning (kmeans)

MBR modeling is based on the assumptions that the input variables are numeric, orthogonal to each other and standardized. The latter two assumptions are taken care of by Principal Components' transformation of raw variables and using the components instead of the raw variables as inputs to MBR. Since the MBR node ignores data points with missing values, in addition to the principal components node, the impute node is also used to replace missing variables using the tree surrogate method. To satisfy the first assumption, the categorical variables are dummy coded. This raises issues by increasing the dimensionality and overfitting to the training data by introducing discontinuity in the response relating inputs and target variables.

Regression Modelling Preparation

Regression models assume that a given dataset is complete, normalized, and contains identifiable linear relationships. To ensure that the dataset contains no missing values, the data is imputed using the tree surrogate method to estimate the missing values. The resulting interval values are then passed through a logarithmic, base 10 function to normalize highly skewed variables. Highly skewed variables for this project are identified as variables with a skewness greater than 1 as identified in the initial Stat Explore node. The correlation investigation from earlier indicates the presence of multiple linear relationships within the dataset which satisfies the third assumption. In order to account for the nominal variables within the dataset, the modelling will be using logit functions instead of simple linear regression models.

Logistic Modelling

The logistic model is able to handle all data types present within the data set because of its use of non-linear log transformations which it can apply to the odds ratio of both interval and class variables.

The logit function is represented as:

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

The logit function is the natural log of the odds that Y equals one of the categories. Since our target variable is continuous, the model will separate the model into bins or depth categories for which to predict the odds.

Model Selection

For each model type, multiple permutations were created in order to find the best setup. In selecting the model, we looked at the Average Squared Error (ASE) for the validation dataset using the model comparison node. As per the scope of this project, the final model will be separated by route type.

Linehaul Model Selection and Comparison

When running the model comparison node on the linehaul dataset, the below output is produced.

Selected Model	*Model Node*	*Model Description*	*Valid: Average Squared Error*	*Train: Average Squared Error*
Y	Reg7	PolyReg Deg2	1.13347	1.19501
	Reg2	PolyReg Deg3	1.14174	1.05528
	Tree2	4 Branch Maximal	1.14887	1.08359
	MBR2	MBR 8	1.14961	0.98839
	MBR3	MBR 4	1.16979	0.76564
	MBR	MBR 16	1.19523	1.15495
	HPNNA	HP Neural	1.21859	1.32531
	Reg	Regression	1.22571	1.31274
	Tree3	Binary maximal	1.22574	1.24186
	Reg4	Linear Stepwise	1.22673	1.31728
	Reg5	Logistic Stepwise AIC Selection	1.22673	1.31728
	Boost	Gradient Boosting Tree	1.22698	1.31019
	Neural	Neural Network	1.24725	1.3824
	DMNeural	DMNeural	1.26498	1.39721
	AutoNeural	AutoNeural	1.44249	1.58904

28 Linehaul Model Selection Results

As we can see, the model that suited our requirements the best was the second degree polynomial regression logit function as it had the lowest ASE on the validation partition of the data. Interestingly though, the selected model was only the 5th best model when comparing the ASE on the training set behind all three permutations of memory based reasoning, the 4 branch maximal decision tree, and the third degree polynomial logit regression models. This is due to those models overfitting to the training data and failing to be representative of the variation that may occur in other sets.
Investigating the second degree polynomial further, we see the definition of the model represented in figure 29.

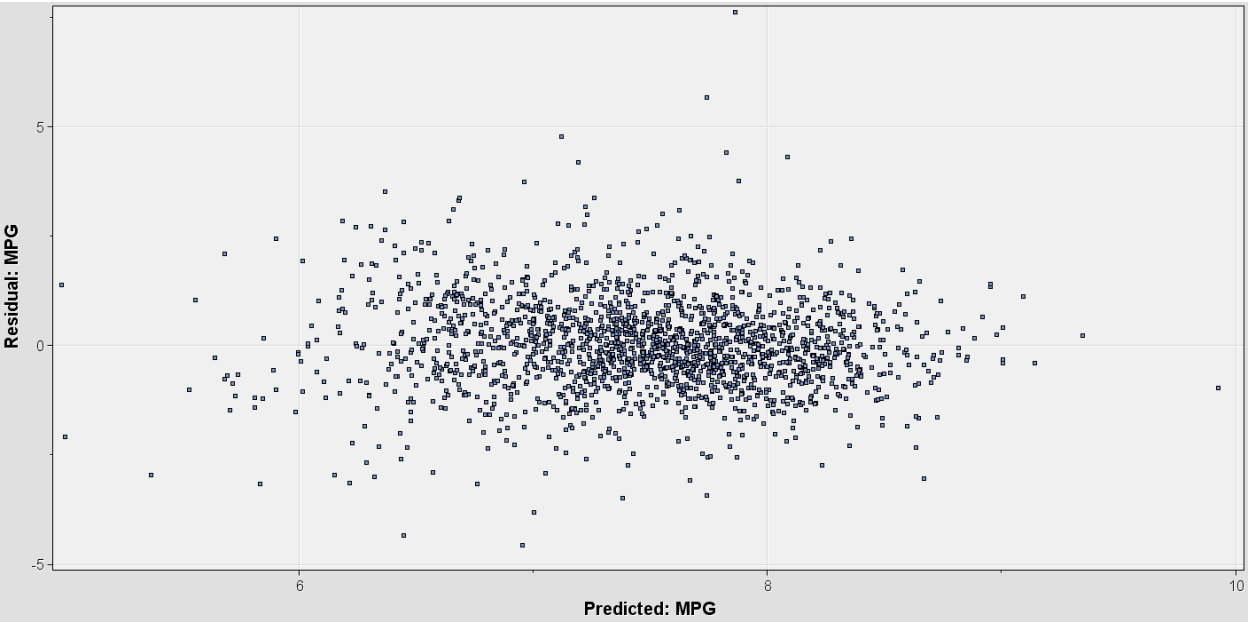
Parameter	Estimate	Standard Error	t Value	Pr> |t|
Intercept	6.1528	0.6238	9.86	<.0001
CRUISE_PERCENT	1.1166	0.2056	5.43	<.0001
DRIVER_FLAGS	-0.017	0.00262	-6.47	<.0001
LG10_BRAKE_EVENTS	0.5051	0.221	2.29	0.0223
LG10_JAKE_BRAKE	0.858	0.0909	9.44	<.0001
LG10_SHORT_IDLE_TIME	-0.6297	0.1119	-5.63	<.0001
REP_DRIVE_AXLE_SET_UP SINGLE AXLE	0.2155	0.0406	5.31	<.0001
REP_ENGINE_MODEL C13	-1.2022	1.193	-1.01	0.3136
REP_ENGINE_MODEL DD15	2.0459	0.5029	4.07	<.0001
REP_ENGINE_MODEL DD15TC	2.1715	0.5282	4.11	<.0001
REP_ENGINE_MODEL ISB	-1.8322	0.7867	-2.33	0.0199
REP_ENGINE_MODEL ISL	-3.1826	0.6531	-4.87	<.0001
REP_ENGINE_MODEL ISX	1.9784	0.5076	3.9	<.0001
REP_ENGINE_MODEL MBE4000	-1.9329	1.1496	-1.68	0.0927
REP_EQUIPMENT_CATEGORY_TYPE_NM DAYCAB	-0.1033	0.0908	-1.14	0.255
REP_EQUIPMENT_CATEGORY_TYPE_NM FI SHELL	0.2746	0.1916	1.43	0.1518
REP_EQUIPMENT_CATEGORY_TYPE_NM SLEEPER	0	.	.	.
REP_MODEL_YEAR 2004	0.7831	0.536	1.46	0.144
REP_MODEL_YEAR 2005	1.4079	0.5519	2.55	0.0108
REP_MODEL_YEAR 2006	2.7839	0.7252	3.84	0.0001
REP_MODEL_YEAR 2007	2.0724	0.8142	2.55	0.0109
REP_MODEL_YEAR 2008	-3.6209	1.0409	-3.48	0.0005
REP_MODEL_YEAR 2009	-1.1725	0.5251	-2.23	0.0256
REP_MODEL_YEAR 2010	-0.769	0.393	-1.96	0.0505
REP_MODEL_YEAR 2011	-0.6728	0.3171	-2.12	0.0339
REP_MODEL_YEAR 2012	-0.4851	0.2332	-2.08	0.0375
REP_MODEL_YEAR 2013	-0.7572	0.1798	-4.21	<.0001
REP_MODEL_YEAR 2014	0.0925	0.1267	0.73	0.4653
REP_MODEL_YEAR 2015	0.0476	0.0665	0.72	0.474
REP_MODEL_YEAR 2016	0	.	.	.
REP_TRANSMISSION_MAKE ALLSN	3.2553	0.6077	5.36	<.0001
REP_TRANSMISSION_MAKE DTDSC	-3.21	0.4918	-6.53	<.0001
REP_TRANSMISSION_MAKE EATON	0	.	.	.
TRUCK_MILAGE	1.42E-06	3.25E-07	4.37	<.0001
CRUISE_PERCENT*IMP_TIME_IN_SERVICE	-0.00038	0.000089	-4.27	<.0001
CRUISE_PERCENT*LG10_BRAKE_EVENTS	-0.407	0.1473	-2.76	0.0058
CRUISE_PERCENT*LG10_JAKE_BRAKE	-0.373	0.0589	-6.33	<.0001
CRUISE_PERCENT*LG10_OVER_RPM	0.2381	0.0643	3.7	0.0002
CRUISE_PERCENT*LG10_OVER_SPEED	-0.3696	0.0411	-8.99	<.0001

CRUISE_PERCENT*LG10_SHORT_IDLE_TIME	0.4747	0.0835	5.69	<.0001
DRIVER_FLAGS*DRIVER_FLAGS	0.000036	7.16E-06	5.07	<.0001
DRIVER_FLAGS*LG10_BRAKE_EVENTS	0.00386	0.00102	3.77	0.0002
DRIVER_FLAGS*LG10_LONG_IDLE_TIME	0.002	0.000647	3.1	0.002
DRIVER_FLAGS*LG10_OVER_RPM	-0.00115	0.000348	-3.32	0.0009
IMP_TIME_IN_SERVICE*IMP_TIME_IN_SERVICE	1.37E-07	5.65E-08	2.43	0.0153
IMP_TIME_IN_SERVICE*LG10_BRAKE_EVENTS	-0.00022	0.000044	-4.98	<.0001
IMP_TIME_IN_SERVICE*LG10_LONG_IDLE_TIME	0.000075	0.000026	2.93	0.0034
IMP_TIME_IN_SERVICE*LG10_OVER_RPM	-0.00012	0.000024	-4.92	<.0001
IMP_TIME_IN_SERVICE*LG10_SHORT_IDLE_TIME	0.000137	0.00004	3.46	0.0005
IMP_TIME_TO_FULL_TIME*LG10_JAKE_BRAKE	0.0586	0.00935	6.27	<.0001
LG10_BRAKE_EVENTS*LG10_JAKE_BRAKE	-0.5374	0.0532	-10.09	<.0001
LG10_BRAKE_EVENTS*LG10_LONG_IDLE_TIME	0.2745	0.0469	5.86	<.0001
LG10_BRAKE_EVENTS*LG10_OVER_RPM	-0.1747	0.0478	-3.65	0.0003
LG10_BRAKE_EVENTS*LG10_SHORT_IDLE_TIME	0.2472	0.0637	3.88	0.0001
LG10_CRUISE_EVENTS*LG10_CRUISE_EVENTS	-0.2429	0.0574	-4.23	<.0001
LG10_CRUISE_EVENTS*LG10_LONG_IDLE_TIME	0.1268	0.0381	3.33	0.0009
LG10_CRUISE_EVENTS*LG10_TOPGEAR_PERCENTAGE	-4.4445	1.9177	-2.32	0.0205
LG10_CRUISE_EVENTS*TRUCK_MILAGE	7.87E-07	1.67E-07	4.71	<.0001
LG10_EXCESS_SPEED*LG10_SEATBELT_TIME	-0.0548	0.0181	-3.03	0.0024
LG10_JAKE_BRAKE*LG10_SHORT_IDLE_TIME	0.1411	0.031	4.55	<.0001
LG10_JAKE_BRAKE*TIME_SINCE_HIRE	-0.012	0.00327	-3.65	0.0003
LG10_LONG_IDLE_TIME*LG10_LONG_IDLE_TIME	-0.2143	0.0285	-7.51	<.0001
LG10_LONG_IDLE_TIME*LG10_OVER_SPEED	0.0667	0.0212	3.15	0.0016
LG10_LONG_IDLE_TIME*LG10_SHORT_IDLE_TIME	-0.1051	0.0255	-4.13	<.0001
LG10_OVER_RPM*LG10_SHORT_IDLE_TIME	0.1855	0.0316	5.88	<.0001
LG10_SHORT_IDLE_TIME*LG10_SHORT_IDLE_TIME	-0.2243	0.032	-7	<.0001
LG10_SHORT_IDLE_TIME*TIME_SINCE_HIRE	0.0147	0.00268	5.47	<.0001
LG10_TOPGEAR_PERCENTAGE*LG10_TOPGEAR_PERCENTAGE	52.3725	5.9051	8.87	<.0001
TIME_SINCE_HIRE*TRUCK_MILAGE	-3.90E-08	1.27E-08	-3.06	0.0022
TRUCK_MILAGE*TRUCK_MILAGE	-1.51E-12	3.10E-13	-4.88	<.0001

29 Second Degree Polynomial Logit Regression Model Definition

The results show that the model is a 67 term polynomial (66 terms above with non-zero estimates plus the intercept) with a Pr > F of <.0001. This indicates this model is very good at accounting for the variance within the validation dataset. However with an R-Square value of .2511, we can say the model has a relatively poor "goodness of fit." This can be explained partially by the variation within the target variable itself as evidenced in the residual plot in figure 30.

Linehaul Model Analysis



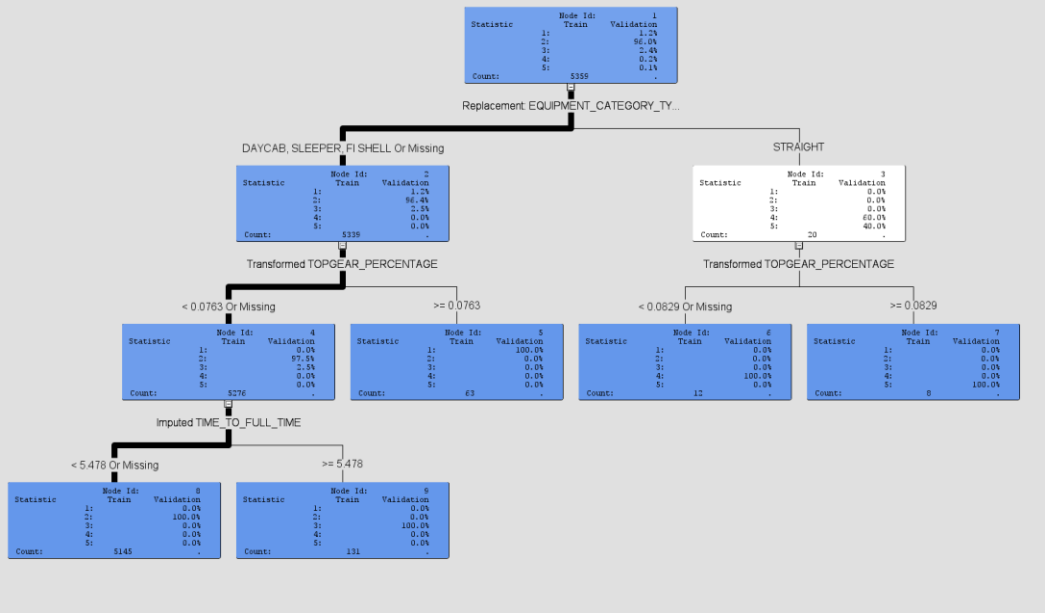
30 Linehaul Model Residual Plot

From this residual plot, it can be observed that the vast majority of data points have residuals resulting in a range of $-2 < y < 2$ though there are multiple data points that fall outside of that range with a couple exceptional negative outliers near the predicted MPG of 7 and two exceptional positive outliers near a predicted value of 8.

These values taken together indicate that in a vacuum the model will perform very well but once outside factors are introduced (weather, altitude, traffic, frequency of hills along route, etc.) the model loses some accuracy. It is recommended that this model be integrated when attempting to estimate MPG for a given route though some additional research into environmental factors be conducted.

Linehaul clustering

To explain some of the variation within the target variable, we perform a cluster analysis by using the selected model to inform a cluster node and subsequently a segment profile node. The cluster node uses the Ward clustering method to identify 5 significant clusters. The data within each of these clusters is identified by the simple tree plot following.



31 Decision Tree for Linehaul Clustering

The rules for each node are given in the output below.

Node = 5
if Transformed TOPGEAR_PERCENTAGE >= 0.07628
AND Replacement: EQUIPMENT_CATEGORY_TYPE_NM IS ONE OF: DAYCAB, SLEEPER, FI SHELL or MISSING
then
Tree Node Identifier = 5
Number of Observations = 63
Predicted: _SEGMENT_=2 = 0.00
Predicted: _SEGMENT_=1 = 1.00
Predicted: _SEGMENT_=3 = 0.00
Predicted: _SEGMENT_=5 = 0.00
Predicted: _SEGMENT_=4 = 0.00

Node = 6
if Transformed TOPGEAR_PERCENTAGE < 0.08289 or MISSING
AND Replacement: EQUIPMENT_CATEGORY_TYPE_NM IS ONE OF: STRAIGHT
then
Tree Node Identifier = 6
Number of Observations = 12
Predicted: _SEGMENT_=2 = 0.00
Predicted: _SEGMENT_=1 = 0.00
Predicted: _SEGMENT_=3 = 0.00

Predicted: _SEGMENT_=5 = 0.00
Predicted: _SEGMENT_=4 = 1.00

Node = 7
if Transformed TOPGEAR_PERCENTAGE >= 0.08289
AND Replacement: EQUIPMENT_CATEGORY_TYPE_NM IS ONE OF: STRAIGHT
then
Tree Node Identifier = 7
Number of Observations = 8
Predicted: _SEGMENT_=2 = 0.00
Predicted: _SEGMENT_=1 = 0.00
Predicted: _SEGMENT_=3 = 0.00
Predicted: _SEGMENT_=5 = 1.00
Predicted: _SEGMENT_=4 = 0.00

Node = 8
if Transformed TOPGEAR_PERCENTAGE < 0.07628 or MISSING
AND Replacement: EQUIPMENT_CATEGORY_TYPE_NM IS ONE OF: DAYCAB, SLEEPER, FI SHELL or MISSING
AND Imputed TIME_TO_FULL_TIME < 5.478 or MISSING
then
Tree Node Identifier = 8
Number of Observations = 5145
Predicted: _SEGMENT_=2 = 1.00
Predicted: _SEGMENT_=1 = 0.00
Predicted: _SEGMENT_=3 = 0.00
Predicted: _SEGMENT_=5 = 0.00
Predicted: _SEGMENT_=4 = 0.00

Node = 9
if Transformed TOPGEAR_PERCENTAGE < 0.07628 or MISSING
AND Replacement: EQUIPMENT_CATEGORY_TYPE_NM IS ONE OF: DAYCAB, SLEEPER, FI SHELL or MISSING
AND Imputed TIME_TO_FULL_TIME >= 5.478
then
Tree Node Identifier = 9
Number of Observations = 131
Predicted: _SEGMENT_=2 = 0.00
Predicted: _SEGMENT_=1 = 0.00
Predicted: _SEGMENT_=3 = 1.00
Predicted: _SEGMENT_=5 = 0.00
Predicted: _SEGMENT_=4 = 0.00

We see that 96% of all data points fall into node 8 where the EQUIPMENT_CATEGORY_TYPE_NM is one of either daycab, sleeper or Fi Shell, the Transformed TOPGEAR_PERCENTAGE is < 0.0763 and the Imputed TIME_TO_FULL_TIME is less than 5.478. This is comforting that the bulk of the data can be placed into the same cluster. All data points involving a straight cab are placed into separate clusters from the largest cluster. Though these data points only account for 0.3% of the data, some additional research may be done into the effectiveness of the straight cab style tractors.

By then using a Segment Profile node, we see more detail about each cluster. By doing so we see that the mean MPG for node 8 is 7.45 whereas the mean MPG for the other clusters is 8.05. This indicates that the datapoints that fall into these smaller clusters may have a positive influence on overall MPG.

Linehaul cluster analysis

We now compare each cluster to one another. By doing so we find the following differences in our target variable.

Segment	*Cluster*	*n*	*Mean*	*Std. Deviation*
1	5	63	7.687095	1.471176
2	8	5145	7.448652	1.256522
3	9	131	8.087069	1.116157
4	6	12	8.78075	1.752665
5	7	8	9.161125	0.423867

After performing an ANOVA analysis on these values, the output below is received.

Segment 1 vs Segment 2: Diff=-0.2384, 95%CI=-0.6729 to 0.1960, p=0.5644
Segment 1 vs Segment 3: Diff=0.4000, 95%CI=-0.1255 to 0.9254, p=0.2305
Segment 1 vs Segment 4: Diff=1.0937, 95%CI=0.0142 to 2.1731, p=0.0453
Segment 1 vs Segment 5: Diff=1.4740, 95%CI=0.1877 to 2.7603, p=0.0152
Segment 2 vs Segment 3: Diff=0.6384, 95%CI=0.3352 to 0.9416, p=0.0000
Segment 2 vs Segment 4: Diff=1.3321, 95%CI=0.3416 to 2.3226, p=0.0022
Segment 2 vs Segment 5: Diff=1.7125, 95%CI=0.4999 to 2.9251, p=0.0011
Segment 3 vs Segment 4: Diff=0.6937, 95%CI=-0.3400 to 1.7273, p=0.3560
Segment 3 vs Segment 5: Diff=1.0741, 95%CI=-0.1741 to 2.3222, p=0.1303
Segment 4 vs Segment 5: Diff=0.3804, 95%CI=-1.1839 to 1.9446, p=0.9642

Several segment combinations (1 & 4, 1 & 5, 2 & 3, 2 & 4, and 2 & 5) have confidence intervals that do not span zero and thus a p value of less than .05. This indicates that even though the number of observations in segments 1, 3, 4 and 5 are limited, the differences in the means of the indicated segments are statistically significant.

By comparing individual segments to one another we can produce a couple conclusions.

First, note that segments 4 and 5 contain only data points using sleeper cabs on their tractors. Every segment that is statistically different from segment 4 is also statistically different from segment 5 and vice versa. The only segment that is not statistically significant from either cluster is segment 3 which has p values with clusters 4 and 5 of .3560 and .1303 respectively. Thus while the differences with segment 3 are not statistically significant, they can still be considered dissimilar. Segments 4 and 5 have a

p value = .9642 thus the two clusters are quite similar and can be treated as one. Taking these facts into consideration, it can be said that with 95% confidence, straight cabs generate dispatches with higher than average MPG and more research should be conducted on the ability to use these tractors on more dispatches.

Another observation that can be made is that segments 2 and 3 are statistically different from one another with segment 3 on average returning .639 MPG better than segment 2. The only difference in these two clusters is that the time to full time for segment 2 is less than 5.5 years whereas the time to full time for segment 3 is greater than 5.5 years. While this seems counter intuitive that those who were promoted faster produce less MPG further investigation is required. Looking further into the mean statistics shows that the average time since hire for segment 2 7.83 years whereas the average time since hire for segment three is a whopping 15.78 years. The average driver in segment 3 has been driving for twice as long as those in segment 2 and thus would be expected to be more experienced and more capable of producing a high MPG. This provides some clarity as to why the split in TIME_TO_FULL_TIME seems counterintuitive.

PND Model Selection and Comparison

When running the model comparison node on the linehaul dataset, the below output is produced.

Selected Model	*Model Node*	*Model Description*	*Valid: Average Squared Error*	*Train: Average Squared Error*
Y	Reg2	PolyReg Deg3	1.38013	1.38832
	Reg7	PolyReg Deg2	1.3892	1.49635
	Tree2	4 Branch Maximal	1.45854	1.52773
	Reg	Regression	1.46228	1.74036
	Reg4	Linear Stepwise	1.46556	1.7469
	Reg5	Logistic Stepwise AIC selection	1.46556	1.7469
	MBR	MBR 16	1.48828	1.51294
	MBR2	MBR 8	1.49499	1.3179
	HPNNA	HP Neural	1.50117	1.70257
	Boost	Gradient Boosting	1.51802	1.74955
	Tree3	Binary maximal	1.54949	1.78369
	MBR3	MBR 4	1.56131	1.06036
	Neural	Neural Network	1.5833	1.83379
	DMNeural	DMNeural	1.60076	1.87283
	AutoNeural	AutoNeural	1.8505	2.15429

32 PND Model Selection Results

In the PND dataset the third degree polynomial model outperforms all other models when comparing via the ASE on the validation partition. As with the linehaul dataset, the memory based reasoning models using 4 and 8 nearest neighbors out-perform in the training partition but are once again over-fitted and are not representative of the data as a whole.

Investigating the third degree polynomial model further we see the following in the output of the node.

Parameter	Estimate	Standard Error	t Value	Pr> |t|
Intercept	8.3211	1.0425	7.98	<.0001
LG10_TOPGEAR_PERCENTAGE	17.5798	2.2602	7.78	<.0001
REP_DRIVE_AXLE_SET_UP SINGLE AXLE	0.2835	0.0762	3.72	0.0002
REP_ENGINE_MODEL C13	3.7838	1.4028	2.7	0.007
REP_ENGINE_MODEL D13	6.186	8.1605	0.76	0.4485

REP_ENGINE_MODEL DD15	-2.6481	1.1071	-2.39	0.0168
REP_ENGINE_MODEL DD15TC	-3.5009	1.1169	-3.13	0.0017
REP_ENGINE_MODEL ISB	-1.8879	1.1759	-1.61	0.1084
REP_ENGINE_MODEL ISL	-3.816	1.1062	-3.45	0.0006
REP_ENGINE_MODEL ISX	-3.0484	1.1056	-2.76	0.0059
REP_ENGINE_MODEL MBE4000	3.5555	1.4014	2.54	0.0112
REP_FIFTH_WHEEL_TYPE FIXED	-0.1769	0.0774	-2.29	0.0223
REP_MODEL C112	-0.5493	0.6355	-0.86	0.3874
REP_MODEL CA125	1.1515	0.2594	4.44	<.0001
REP_MODEL CL112	-2.3571	0.6399	-3.68	0.0002
REP_MODEL CL120	1.9086	0.6792	2.81	0.005
REP_MODEL M2106	0	.	.	.
REP_MODEL VNL42T	0	.	.	.
REP_MODEL VNL42T30	0.593	0.2053	2.89	0.0039
REP_MODEL VNL42T67	0	.	.	.
REP_MODEL_YEAR 2004	-1.1622	0.7993	-1.45	0.146
REP_MODEL_YEAR 2005	-0.7082	0.7572	-0.94	0.3497
REP_MODEL_YEAR 2006	-1.1911	0.7499	-1.59	0.1123
REP_MODEL_YEAR 2007	-1.2669	0.7528	-1.68	0.0925
REP_MODEL_YEAR 2008	0.5026	0.6221	0.81	0.4191
REP_MODEL_YEAR 2009	1.1652	0.3938	2.96	0.0031
REP_MODEL_YEAR 2010	0.5322	0.3729	1.43	0.1536
REP_MODEL_YEAR 2011	-0.1346	0.3856	-0.35	0.727
REP_MODEL_YEAR 2012	0.2644	0.3776	0.7	0.484
REP_MODEL_YEAR 2013	0.2859	0.3778	0.76	0.4493
REP_MODEL_YEAR 2014	0.4495	0.3824	1.18	0.2398
REP_MODEL_YEAR 2015	0.2836	0.3836	0.74	0.4597
REP_MODEL_YEAR 2016	0.497	0.3893	1.28	0.2018
REP_TRANSMISSION_MAKE ALLSN	0	.	.	.
REP_TRANSMISSION_MAKE DTDSC	-5.7494	1.1091	-5.18	<.0001
REP_TRANSMISSION_MAKE EATON	0.3869	0.1785	2.17	0.0303
REP_TRANSMISSION_MAKE MERTR	0	.	.	.
CRUISE_PERCENT*TRUCK_MILAGE	7.16E-06	7.82E-07	9.16	<.0001
DRIVER_FLAGS*IMP_TIME_IN_SERVICE	4.30E-06	1.06E-06	4.05	<.0001
DRIVER_FLAGS*LG10_SHORT_IDLE_TIME	-4.73E-03	0.00108	-4.38	<.0001
LG10_BRAKE_EVENTS*LG10_BRAKE_EVENTS	0.5543	1.19E-01	4.66	<.0001
LG10_CRUISE_EVENTS*LG10_JAKE_BRAKE	-0.8375	0.1537	-5.45	<.0001
LG10_JAKE_BRAKE*TRUCK_MILAGE	1.65E-06	3.00E-07	5.51	<.0001
LG10_SEATBELT_TIME*TIME_SINCE_HIRE	-4.06E-02	0.00734	-5.53	<.0001
LG10_TOPGEAR_PERCENTAGE*TIME_SINCE_HIRE	-8.98E-01	2.50E-01	-3.59	0.0003
CRUISE_PERCENT*DRIVER_FLAGS*DRIVER_FLAGS	0.000023	8.90E-06	2.63	0.0086

CRUISE_PERCENT*DRIVER_FLAGS*LG10_CRUISE_EVENTS	-0.0118	0.00214	-5.52	<.0001
CRUISE_PERCENT*DRIVER_FLAGS*LG10_OVER_SPEED	-0.00589	0.00111	-5.3	<.0001
CRUISE_PERCENT*IMP_TIME_IN_SERVICE*LG10_JAKE_BRAKE	0.000469	0.000222	2.11	0.0347
CRUISE_PERCENT*LG10_BRAKE_EVENTS*LG10_EXCESS_SPEED	0.4551	0.1014	4.49	<.0001
CRUISE_PERCENT*LG10_BRAKE_EVENTS*TRUCK_MILAGE	-7.23E-07	3.15E-07	-2.29	0.0218
CRUISE_PERCENT*LG10_JAKE_BRAKE*LG10_JAKE_BRAKE	2.87E-01	0.0924	3.11	0.0019
CRUISE_PERCENT*LG10_JAKE_BRAKE*TRUCK_MILAGE	-2.63E-06	4.01E-07	-6.56	<.0001
CRUISE_PERCENT*LG10_SEATBELT_TIME*LG10_SEATBELT_TIME	-0.0456	0.0191	-2.39	0.017
CRUISE_PERCENT*TRUCK_MILAGE*TRUCK_MILAGE	-6.24E-12	7.35E-13	-8.49	<.0001
DRIVER_FLAGS*IMP_TIME_IN_SERVICE*LG10_BRAKE_EVENTS	-6.09E-06	6.67E-07	-9.13	<.0001
DRIVER_FLAGS*IMP_TIME_IN_SERVICE*LG10_SHORT_IDLE_TIME	1.97E-06	4.47E-07	4.4	<.0001
DRIVER_FLAGS*LG10_BRAKE_EVENTS*LG10_SEATBELT_TIME	-3.18E-03	4.59E-04	-6.92	<.0001
DRIVER_FLAGS*LG10_BRAKE_EVENTS*LG10_SHORT_IDLE_TIME	0.00429	0.000588	7.29	<.0001
DRIVER_FLAGS*LG10_BRAKE_EVENTS*TRUCK_MILAGE	4.73E-09	1.22E-09	3.86	0.0001
DRIVER_FLAGS*LG10_CRUISE_EVENTS*LG10_SHORT_IDLE_TIME	0.00146	0.000659	2.22	0.0265
DRIVER_FLAGS*LG10_OVER_RPM*LG10_OVER_SPEED	0.000688	0.000231	2.97	0.003
DRIVER_FLAGS*LG10_SEATBELT_TIME*LG10_SHORT_IDLE_TIME	1.03E-03	0.000309	3.35	0.0008
DRIVER_FLAGS*LG10_SEATBELT_TIME*TIME_SINCE_HIRE	0.000299	0.000048	6.19	<.0001
DRIVER_FLAGS*LG10_SHORT_IDLE_TIME*LG10_SHORT_IDLE_TIME	-0.00198	0.000427	-4.64	<.0001
IMP_TIME_IN_SERVICE*IMP_TIME_IN_SERVICE*LG10_LONG_IDLE_TIME	3.38E-08	1.31E-08	2.57	0.0101
IMP_TIME_IN_SERVICE*LG10_CRUISE_EVENTS*LG10_JAKE_BRAKE	0.00033	0.000129	2.57	0.0103
IMP_TIME_IN_SERVICE*LG10_LONG_IDLE_TIME*LG10_LONG_IDLE_TIME	-0.00005	0.000023	-2	0.0455
IMP_TIME_IN_SERVICE*LG10_OVER_RPM*LG10_SEATBELT_TIME	1.36E-04	2.20E-05	6.25	<.0001
IMP_TIME_IN_SERVICE*LG10_SHORT_IDLE_TIME*LG10_SHORT_IDLE_TIME	-3.00E-05	8.64E-06	-3.87	0.0001
IMP_TIME_TO_FULL_TIME*LG10_EXCESS_SPEED*LG10_LONG_IDLE_TIME	-8.58E-02	0.0305	-2.82	0.0049
IMP_TIME_TO_FULL_TIME*LG10_OVER_SPEED*TIME_SINCE_HIRE	-0.00509	0.00173	-2.94	0.0033
IMP_TIME_TO_FULL_TIME*TIME_SINCE_HIRE*TIME_SINCE_HIRE	0.000707	0.0001	7.08	<.0001
LG10_BRAKE_EVENTS*LG10_BRAKE_EVENTS*LG10_BRAKE_EVENTS	-0.1828	0.0372	-4.91	<.0001
LG10_BRAKE_EVENTS*LG10_BRAKE_EVENTS*LG10_CRUISE_EVENTS	-2.50E-01	0.0453	-5.51	<.0001
LG10_BRAKE_EVENTS*LG10_BRAKE_EVENTS*TIME_SINCE_HIRE	0.0127	0.00284	4.47	<.0001
LG10_BRAKE_EVENTS*LG10_BRAKE_EVENTS*TRUCK_MILAGE	3.86E-07	5.89E-08	6.56	<.0001
LG10_BRAKE_EVENTS*LG10_JAKE_BRAKE*TRUCK_MILAGE	-7.45E-07	1.64E-07	-4.55	<.0001
LG10_BRAKE_EVENTS*LG10_LONG_IDLE_TIME*LG10_LONG_IDLE_TIME	0.1139	0.0153	7.45	<.0001
LG10_BRAKE_EVENTS*LG10_OVER_RPM*TRUCK_MILAGE	-6.74E-07	9.18E-08	-7.34	<.0001
LG10_CRUISE_EVENTS*LG10_CRUISE_EVENTS*LG10_LONG_IDLE_TIME	-3.55E-01	0.1208	-2.94	0.0033
LG10_CRUISE_EVENTS*LG10_JAKE_BRAKE*LG10_OVER_RPM	-0.1304	0.044	-2.96	0.0031
LG10_CRUISE_EVENTS*LG10_JAKE_BRAKE*LG10_OVER_SPEED	0.2925	0.0565	5.18	<.0001
LG10_CRUISE_EVENTS*LG10_LONG_IDLE_TIME*LG10_LONG_IDLE_TIME	0.3807	0.0525	7.25	<.0001
LG10_CRUISE_EVENTS*LG10_LONG_IDLE_TIME*LG10_OVER_SPEED	-0.1878	0.0513	-3.66	0.0003
LG10_CRUISE_EVENTS*LG10_LONG_IDLE_TIME*LG10_SHORT_IDLE_TIME	-0.1949	0.0412	-4.73	<.0001
LG10_CRUISE_EVENTS*LG10_OVER_SPEED*LG10_SHORT_IDLE_TIME	-1.08E-01	0.0336	-3.2	0.0014
LG10_CRUISE_EVENTS*LG10_SHORT_IDLE_TIME*LG10_SHORT_IDLE_TIME	0.206	0.0241	8.56	<.0001

LG10_JAKE_BRAKE*LG10_OVER_SPEED*LG10_SHORT_IDLE_TIME	-5.74E-02	0.0149	-3.85	0.0001
LG10_JAKE_BRAKE*LG10_SHORT_IDLE_TIME*LG10_SHORT_IDLE_TIME	0.0378	0.00728	5.19	<.0001
LG10_JAKE_BRAKE*TIME_SINCE_HIRE*TRUCK_MILAGE	-4.04E-08	8.78E-09	-4.6	<.0001
LG10_LONG_IDLE_TIME*LG10_LONG_IDLE_TIME*LG10_LONG_IDLE_TIME	-8.89E-02	0.00991	-8.97	<.0001
LG10_LONG_IDLE_TIME*LG10_OVER_RPM*LG10_OVER_SPEED	5.16E-02	0.0206	2.5	0.0123
LG10_LONG_IDLE_TIME*LG10_OVER_SPEED*LG10_SHORT_IDLE_TIME	5.72E-02	0.019	3.01	0.0026
LG10_OVER_RPM*LG10_OVER_RPM*LG10_OVER_RPM	-0.0998	0.0155	-6.43	<.0001
LG10_OVER_RPM*LG10_OVER_RPM*LG10_SEATBELT_TIME	-0.0743	0.0151	-4.9	<.0001
LG10_OVER_RPM*LG10_OVER_RPM*LG10_SHORT_IDLE_TIME	0.103	0.0155	6.66	<.0001
LG10_OVER_RPM*TIME_SINCE_HIRE*TIME_SINCE_HIRE	-0.00042	0.000136	-3.08	0.0021
LG10_OVER_RPM*TRUCK_MILAGE*TRUCK_MILAGE	1.18E-12	1.98E-13	5.98	<.0001
LG10_SEATBELT_TIME*LG10_SEATBELT_TIME*LG10_SEATBELT_TIME	0.0463	0.00571	8.12	<.0001
LG10_SEATBELT_TIME*LG10_SHORT_IDLE_TIME*TRUCK_MILAGE	-2.56E-07	4.58E-08	-5.6	<.0001
LG10_SEATBELT_TIME*TIME_SINCE_HIRE*TIME_SINCE_HIRE	0.000432	0.000162	2.67	0.0076
LG10_SHORT_IDLE_TIME*LG10_SHORT_IDLE_TIME*LG10_SHORT_IDLE_TIME	-0.032	0.00672	-4.75	<.0001
LG10_SHORT_IDLE_TIME*LG10_SHORT_IDLE_TIME*TIME_SINCE_HIRE	-0.00365	0.00125	-2.92	0.0035
LG10_TOPGEAR_PERCENTAGE*TIME_SINCE_HIRE*TRUCK_MILAGE	0.000013	3.47E-06	3.81	0.0001
LG10_TOPGEAR_PERCENTAGE*TRUCK_MILAGE*TRUCK_MILAGE	-6.49E-10	3.09E-10	-2.1	0.0361
TRUCK_MILAGE*TRUCK_MILAGE*TRUCK_MILAGE	-1.68E-18	2.48E-19	-6.78	<.0001

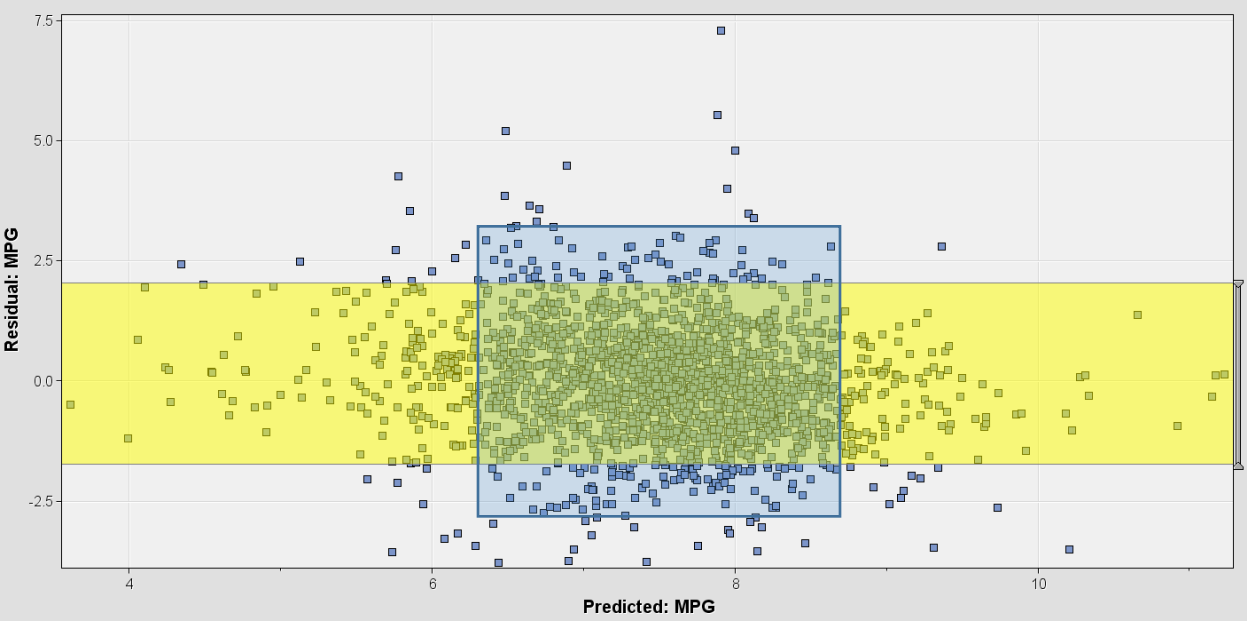
33 Third Degree Polynomial Logit Regression Definition

A few conclusions from the production of this model is that for every increase of .01 in the percent of miles spent in the top gear, the MPG variable is expected to increase by .17. This is the largest difference of all predictor variables produced by this model. Additionally the year model with the largest negative impact on overall MPG is if the tractor was produced in 2007. Conversely, the best year model indicated by this model is the 2009 model year.

PND model analysis

As with the linehaul model, we see a model with a Pr >F <.0001 but with a relatively poor R-Square value of .3686, thus giving us a model that does a good job of accounting for variance within the data but with a relatively poor “goodness of fit” score due to environmental factors outside of the control of this study.

To analyze the effectiveness of each term, the estimate column is used.

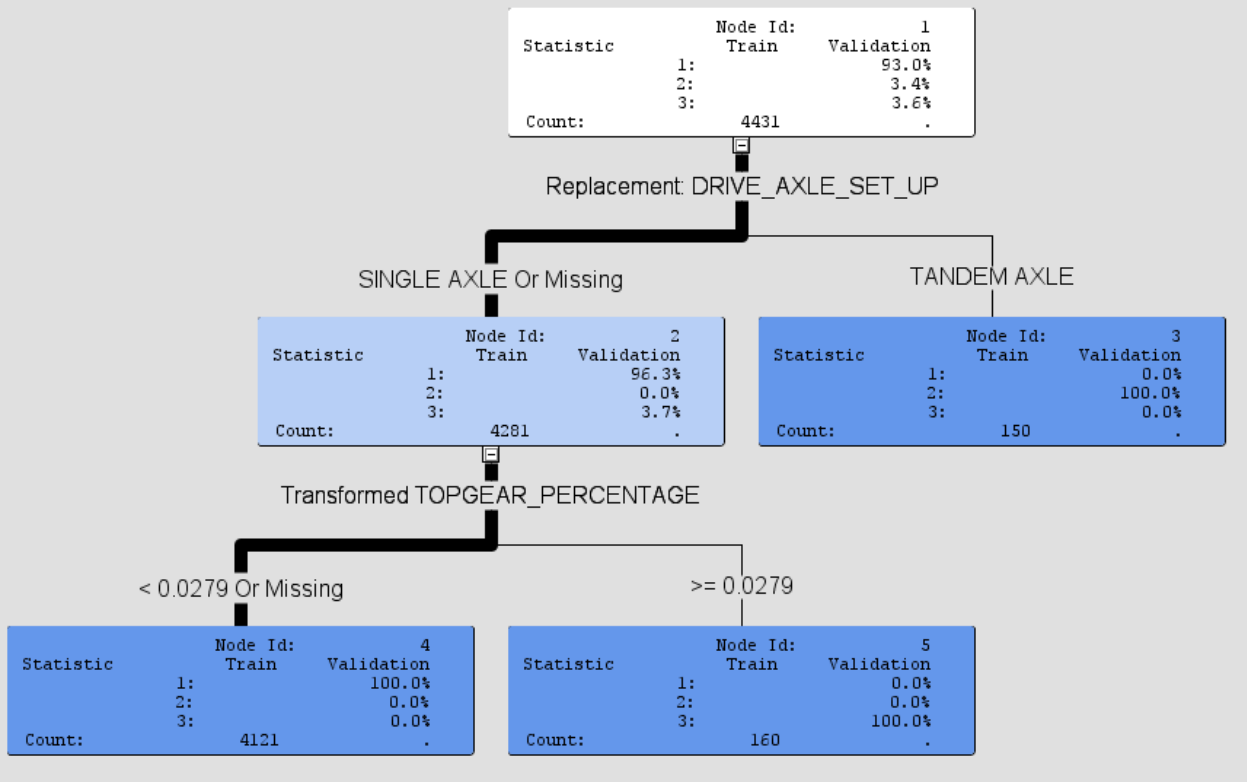


34 PND Selected Model Residual Plot

Viewing the residual plot for this model shows that data points consistently retain a residual value between -1.7 < y < 1.7 (yellow highlighted area) for all predicted values. In the predicted range of 6.2 < x < 8.7 this residual range is extended slightly to -2.7 < y < 2.7 (blue highlighted area). This sort of consistency in a residual plot is expected with roughly normally distributed data with high variance though upon implementation the confidence interval may need to be adjusted depending on the predicted output.

PND clustering

Again to gain further insight into the efficacy of the model we produced, we use it to inform a clustering analysis. The results of the analysis are contained within the following tree.



35 Decision Tree for PND Clusters

The PND dataset is quickly broken into three nodes that follow the following set of rules.

Node = 3
if Replacement: DRIVE_AXLE_SET_UP IS ONE OF: TANDEM AXLE
then
Tree Node Identifier = 3
Number of Observations = 150
Predicted: _SEGMENT_=1 = 0.00
Predicted: _SEGMENT_=3 = 0.00
Predicted: _SEGMENT_=2 = 1.00

Node = 4
if Transformed TOPGEAR_PERCENTAGE < 0.02793 or MISSING
AND Replacement: DRIVE_AXLE_SET_UP IS ONE OF: SINGLE AXLE or MISSING
then
Tree Node Identifier = 4
Number of Observations = 4121
Predicted: _SEGMENT_=1 = 1.00
Predicted: _SEGMENT_=3 = 0.00
Predicted: _SEGMENT_=2 = 0.00

Node = 5

if Transformed TOPGEAR_PERCENTAGE >= 0.02793
AND Replacement: DRIVE_AXLE_SET_UP IS ONE OF: SINGLE AXLE or MISSING
then
Tree Node Identifier = 5
Number of Observations = 160
Predicted: _SEGMENT_=1 = 0.00
Predicted: _SEGMENT_=3 = 1.00
Predicted: _SEGMENT_=2 = 0.00

We see that the clustering analysis separates all dispatches using a tractor with a tandem axle (a total of 150 observations) from the observations with a single axle. Single Axle observations are then split along the TOPGEAR_PERCENTAGE variable at a value of .02793 or when the tractor is using the top most gear for 2.7% of the distance driven in the total dispatch.

PND clustering analysis

Through using the Summary Statistics table in the segment profile node we find the following values

Segment	*Node*	*n*	*Mean*	*Std. Deviation*
1	4	4121	7.485748	1.45689
2	3	150	7.139193	1.479185
3	5	160	7.460413	2.029166

By once again performing an ANOVA analysis on these values the following output is given.

Segment 1 vs Segment 2: Diff=-0.3466, 95%CI=-0.6353 to -0.0578, p=0.0136
Segment 1 vs Segment 3: Diff=-0.0253, 95%CI=-0.3052 to 0.2545, p=0.9754
Segment 2 vs Segment 3: Diff=0.3212, 95%CI=-0.0735 to 0.7160, p=0.1367

This analysis shows that the only Segments that have a significant difference is the difference between segments 1 and 2. Segment one are those which used a single axle and have a low TOPGEAR_PERCENTAGE value and segment two are those which use a tandem axle tractor. Tractors with a tandem axle setup should be expected to have a lower MPG as the additional axle mechanically lowers tractor efficiency and those tractors are typically hauling heavier loads. Further investigations should be performed in an attempt to minimize the use of tandem axle trailers in PND routes to only hauls where they are necessary as they return a statistically significantly lower MPG.

Conclusions

Throughout our study, we identified patterns within the packet data pulled from routes run during the June 2017 timeframe. Patterns identified within the data were consistent with you hypotheses posited prior to beginning the study. As drivers become more experienced, they are more likely to produce a high MPG due to their experience, newer trucks with lower mileage are more likely to produce a high MPG than their older, more worn down counter parts, and the more often a driver is able to utilize the cruise control and keep the tractor in the top most gear the higher their resulting MPG will be.

We then used these patterns to produce a predictive model for the MPG of all routes run separated by dispatch type. For both dispatch types, a logistic model proved to return the most accurate results when comparing the average squared error of each model. For the linehaul dispatch, a second degree polynomial logistic model was deemed the most accurate. The PND dispatch types are best predicted by a cubic logistic model.

Lastly we used clustering to identify discrete groups of similar dispatches with similar characteristics.

From the linehaul dispatches, we identified 5 clusters. Those can be identified as the following: those that employed a straight cab tractor with high utilization of the top gear, those that employed a straight cab with low utilization of the top gear, non-straight cabs with high top gear percentage, non-straight cabs with low top gear percentage and drivers who were quickly hired to full time, and non-straight cabs with low top gear percentage and drivers who were slowly hired to full time.

Of the PND dispatches, 3 clusters were identified. Those are hauls using tractors with tandem axles, hauls using single axles with low top gear percentage, and hauls using single axles with high top gear percentage.

We then identified which clusters were significantly different from one another, why those differences are significant, and suggested further actions to be taken. Conclusions from the cluster analysis can be summarized as the company should further investigate how to minimize the use of tandem axle tractors, maximize the use of straight cab tractors and hire drivers with more experience in order to optimize company-wide fuel usage.

Appendices

Appendix A – Glossary of Terms

Less than Truck Load – A method of shipping in which multiple shippers are allowed to share space on the same truck. Packages are typically between 150 and 15,000 lbs. and are usually palletized.

Linehaul – A type of route where the point of origin and termination are two different service centers. ODFL utilizes line haul routes to carry customer packages across long distances, transferring goods from one service center to another.

Packet – A set of data collected from the sensors on the truck. This is quantitative and grouped by the time from when a truck is turned on to when it is shut off

Pickup and Delivery (PND) – A type of route where the point of origin and termination are the same service center. A driver on PND routes will be travelling to customer locations to either pickup goods that have been scheduled for shipment to bring back to the service center or taking goods from a service center to the customer location for final delivery.

Service Center – The Old Dominion Freight Line facility at which customer packages are stored until they are loaded onto a trailer and shipped to their destination

Tractor – The truck part of an “18-wheeler”. Tractors come in two base variations: Sleepers and Day cabs. Day cabs only have one row of seats and are generally only used on routes that would require less than a day to complete. Sleepers have a place where the driver can rest behind the driver and passenger seats and are used on longer routes where the driver may be required to spend the night on the road.

Appendix B – Variable Explanation

	Variable Name	Variable Description
1	DISPATCH_NBR	identification number given to each individual dispatch
2	DISPATCH_BEGIN_TIME	Time at which the Dispatch is began
3	DISPATCH_END_TIME	Time at which the Dispatch is marked complete
4	PACKET_START_TIME	First time the tractor is turned on
5	PACKET_END_TIME	Last time the tractor is turned off
6	PACKET_ID	Identification number given to each packet of truck sensor data
7	DRIVER_LOGIN	Login number for the driver
8	TRUCK_NUMBER	Truck identification number
9	LONG_IDLE_THRESHOLD	Threshold of how long a truck may idle before being marked as a long idle. An idle occurrence shorter than this threshold is considered a short idle
10	START_DATE	Date of the Packet
11	RPM_THRESHOLD	Threshold dictating what RPM limit under which the driver should stay
12	OVER_SPEED_THRESHOLD	Threshold dictating what speed the driver should stay under. This threshold is lower than the excess speed threshold and should be considered a "warning"
13	EXCESS_SPEED_THRESHOLD	Threshold dictating what speed the driver should stay under. This threshold is higher and thus a more critical warning than over speed threshold
14	TRUCK_MILAGE	Number of miles on the truck at the beginning of the data packet
15	PACKET_MILES	Number of miles travelled during the duration of the data packet
16	OVER_RPM	Number of times the truck exceeded the rpm threshold
17	OVER_SPEED	Number of times the truck exceeded the over speed threshold
18	EXCESS_SPEED	number of times the truck exceeded the excess speed threshold
19	LONG_IDLE_TIME	Amount of time the truck was flagged as having spent in the long idle state
20	LONG_IDLE_COUNT	number of long idle occurrences in a data packet
21	SHORT_IDLE_TIME	Amount of time the truck was flagged as having spent in the short idle time state
22	SHORT_IDLE_COUNT	Number of short idle occurrences in a data packet

23	GAL_FUEL	Number of gallons of fuel used during a data packet
24	MPG	Average Miles Per Gallon of Fuel Used by Dispatch Number
24	CRUISE_EVENTS	Number of times the cruise control was engaged during a packet
25	CRUISE_TIME	Amount of time spent with the cruise control engaged
26	CRUISE_FUEL	Amount of fuel used while the cruise control was engaged
27	CRUISE_DISTANCE	Distance travelled while the cruise control was engaged
28	TOPGEAR_TIME	Amount of time spent in the tractor's highest gear
29	TOPGEAR_DISTANCE	Distance travelled while the tractor was in its highest gear
30	SEATBELT_TIME	Amount of time the seatbelt was used during a packet
31	BRAKE_EVENTS	Number of times the driver braked hard
32	SPEED_GOV_SETTING	Highest speed the truck will go on flat ground
33	JAKE_BRAKE	Number of times the jake brake was used
34	DRIVER_FLAGS	Total number of flags incurred by the driver in their history
35	ASSET_NBR	Identification number given to each asset owned by Old Dominion Freight Line
36	ASSET_DESCR	Brief description of every asset owned by Old Dominion Freight Line
37	EQUIPMENT_CATEGORY_TYPE_NM	Type of cab each tractor has. Day cab or Sleeper. Sleepers have additional space for the driver to sleep in on long hauls
38	ESTIMATED_WEIGHT_OF_FUEL	Estimate of the additional weight incurred by the amount of fuel present in the truck
39	TIME_IN_SERVICE	Amount of time a tractor has been in service for ODFL
40	IN_SERVICE_DTE	Date of when the tractor entered service for ODFL
41	MAKE	Manufacturer of the Tractor
42	MODEL	Tractor model number
43	MODEL_YEAR	Year the tractor was manufactured
44	FIFTH_WHEEL_TYPE	Indicates if the fifth wheel on the tractor is fixed or if it can slide as needed
45	LICENSE_WEIGHT	Weight the tractor is licensed for
46	MAINTENANCE_STATUS_TXT	indicates the current status of the tractor
47	ENGINE_MAKE_TXT	Make of the engine in the tractor

48	ENGINE_MODEL	Model of the engine in the tractor
49	TRANSMISSION_MAKE	Make of the transmission in the tractor
50	DECKS_FLG	Flag to indicate if the asset has a deck
51	LIFTGATE_FLG	Flag to indicate if the asset has a lift gate
52	SKIRT_FLG	Flag to indicate if the asset has a skirt
53	THERMALERT_FLG	Indicates if the tractor was over temperature during the packet
54	TIRE_PSI_FLG	Indicates a low tire pressure warning
55	CAB_TYPE	Type of cab each tractor has. Day cab or Sleeper. Sleepers have additional space for the driver to sleep in on long hauls
56	DRIVE_AXLE_SET_UP	Indicates if the tractor has a single drive axle or if there are multiple axles working in tandem
57	GOVERNED_SPEED	Highest speed the truck will go under its own power
58	METER_READING	Odometer reading of tractor at its last inspection
59	METER_READING_DTE	Date of last odometer reading
60	REAR_AXLE_RATIO	Ratio of the rear axle
61	SERVICE_CENTER_KEY	Indicates the service center that the driver is based at
62	POSITION_DESCR	Brief description of the position maintained by the driver
63	FULL_TM_FLG	Flag to indicate if the driver is full time or part time
64	HOURLY_SALARY_FLG	Flag to indicate if the driver is a hourly or salary employee
65	ORIGINAL_HIRE_DTE	Original hire date of the driver
66	TIME_SINCE_HIRE	Number of years since the original hire date of the driver
67	FULL_TIME_HIRE_DTE	Date at which the driver was hired as a full time employee
68	TIME_TO_FULL_TIME	Number of years between when the driver was originally hired and when they were made full time
69	ELOG_TRAINED_FLG	Indicates if the driver completed the ELOG training
70	ELOG_CERTIFIED_FLG	Indicates if the driver is ELOG certified
71	LINEHAUL_PACOS_TRAINED	Indicates if the driver completed the Linehaul PACOS training
72	PND_PACOS_TRAINED	indicates if the driver completed the PND PACOS training

Appendix C – Database Query for finalized dataset

```
SELECT
  T.DISPATCH_NBR,
  MIN(T.DISPATCH_BEGIN_TM) as DISPATCH_BEGIN_TIME,
  MAX(T.DISPATCH_END_TM)as DISPATCH_END_TIME,
  MIN(P.STRT_DATIME)AS PACKET_START_TIME,
  MAX(P.END_DATIME) as PACKET_END_TIME,
  AVG(P.PACKETID) as PACKET_ID,
  AVG(P.LOGIN) AS DRIVER_LOGIN,
  AVG(P.VEHICLE_NUMBER) as TRUCK_NUMBER,
  AVG(P.LONG_IDLE_THRESH) as LONG_IDLE_THRESHOLD,
  SUBSTR(TO_CHAR(P.STRT_DATIME),1,9) AS START_DATE ,
  AVG(P.RPM_THRESH)as RPM_THRESHOLD,
  AVG(P.OVER_SPEED_THRESH)as OVER_SPEED_THRESHOLD,
  AVG(P.XS_SPEED_THRESH)as EXCESS_SPEED_THRESHOLD,
  MIN(P.STRT_ODOM) as TRUCK_MILAGE,
  SUM(P.TRAVELED_MILES) as PACKET_MILES,
  SUM(P.OVER_RPM)as OVER_RPM,
  SUM(P.OVER_SPD) as OVER_SPEED ,
  SUM(P.XS_SPD) as EXCESS_SPEED ,
  SUM(P.LONG_IDLE_TIME)as LONG_IDLE_TIME,
  SUM(P.LONG_IDLE_CNT) as LONG_IDLE_COUNT,
  SUM(P.SHORT_IDLE_TIME)as SHORT_IDLE_TIME,
  SUM(P.SHORT_IDLE_CNT) as SHORT_IDLE_COUNT,
  SUM(P.GAL_FUEL)AS GAL_FUEL,
  AVG(P.MPG) as MPG,
  SUM(P.CRUISE_EVENTS)as CRUISE_EVENTS,
  SUM(P.CRUISE_TIME)as CRUISE_TIME,
  SUM(P.CRUISE_FUEL)as CRUISE_FUEL,
  SUM(P.CRUISE_DIST)as CRUISE_DISTANCE,
  SUM(P.TOPGEAR_TIME)as TOPGEAR_TIME,
  SUM(P.TOPGEAR_DIST)as TOPGEAR_DISTANCE,
  SUM(P.SEATBELT_TIME)as SEATBELT_TIME,
  SUM(P.BRAKE_EVENTS)as BRAKE_EVENTS,
  AVG(P.SPEED_GOV_SETTING)as SPEED_GOV_SETTING,
  SUM(P.ENGINE_JACOB_BRAKE) as JAKE_BRAKE,
  AVG(P.DRIVER_FLAGS)as DRIVER_FLAGS,
  AVG(A.ASSET_NBR)as ASSET_NBR,
  A.ASSET_DESCR,
  A.EQUIPMENT_CATEGORY_TYPE_NM,
```

```
  AVG(A.ESTIMATED_WEIGHT_OF_FUEL)as ESTIMATED_WEIGHT_OF_FUEL,
  round(TO_NUMBER(SYSDATE - A.IN_SERVICE_DTE)) as TIME_IN_SERVICE,
  A.IN_SERVICE_DTE,
  A.ACTIVITY_CD_KEY,
  A.MAKE,
  A.MODEL,
  A.MODEL_YEAR,
  A.FIFTH_WHEEL_TYPE,
  A.LICENSE_WEIGHT,
  A.MAINTENANCE_STATUS_TXT,
  A.ENGINE_MAKE_TXT,
  A.ENGINE_MODEL,
  A.TRANSMISSION_MAKE,
  A.DECKS_FLG,
  A.LIFTGATE_FLG,
  A.SKIRT_FLG,
  A.THERMALERT_FLG,
  A.TIRE_PSI_FLG,
  A.CAB_TYPE,
  A.DRIVE_AXLE_SET_UP,
  A.GOVERNED_SPEED,
  A.METER_READING,
  A.METER_READING_DTE,
  A.REAR_AXLE_RATIO,
  E.SERVICE_CENTER_KEY,
  E.POSITION_DESCR,
  E.FULL_TM_FLG,
  E.HOURLY_SALARY_FLG,
  E.ORIGINAL_HIRE_DTE,
  round(TO_NUMBER((SYSDATE - E.ORIGINAL_HIRE_DTE)/365), 3) as
TIME_SINCE_HIRE,
  E.FULL_TIME_HIRE_DTE,
  round(TO_NUMBER((E.FULL_TIME_HIRE_DTE - E.ORIGINAL_HIRE_DTE)/365), 3)
as TIME_TO_FULL_TIME,
  D.ELOG_TRAINED_FLG,
  D.ELOG_CERTIFIED_FLG,
  D.LINEHAUL_PACOS_TRAINED,
  D.PND_PACOS_TRAINED

FROM SHARED_ADMIN.SHA_DISPATCHBYTRACTOR_A T
```

```
JOIN SHARED_ADMIN.SHA_DISPATCHDRIVER_A R on R.DISPATCH_NBR =
T.DISPATCH_NBR
JOIN SHARED_ADMIN.SHA_DRIVER_D D on R.DRIVER_KEY = D.DRIVER_KEY --
321778
LEFT JOIN SHARED_ADMIN.SHA_EMPLOYEE_D E
  on D.DRIVER_EMPLOYEE_KEY = E.EMPLOYEE_KEY
JOIN SHARED_ADMIN.PERFORMXBYDRIVERDATA P on to_char(P.Login) =
to_char(D.DRIVER_EMPLOYEE_KEY)
       and (
        T.DISPATCH_BEGIN_TM between P.STRT_DATIME and P.END_DATIME
          or
        P.STRT_DATIME between T.DISPATCH_BEGIN_TM and T.DISPATCH_END_TM
     )
LEFT JOIN SHARED_ADMIN.SHA_ASSET_D A
  on P.VEHICLE_NUMBER = A.ASSET_NBR
WHERE T.DISPATCH_BEGIN_TM like '%JUN-17%'
GROUP by
  T.DISPATCH_NBR, P.STRT_DATIME, A.ASSET_DESCR,
 A.EQUIPMENT_CATEGORY_TYPE_NM, A.IN_SERVICE_DTE,
 A.ACTIVITY_CD_KEY, A.MAKE, A.MODEL, A.MODEL_YEAR,
 A.FIFTH_WHEEL_TYPE, A.LICENSE_WEIGHT, A.MAINTENANCE_STATUS_TXT,
 A.ENGINE_MAKE_TXT, A.ENGINE_MODEL, A.TRANSMISSION_MAKE,
 A.DECKS_FLG, A.LIFTGATE_FLG, A.SKIRT_FLG, A.THERMALERT_FLG,
 A.TIRE_PSI_FLG, A.CAB_TYPE, A.DRIVE_AXLE_SET_UP, A.GOVERNED_SPEED,
 A.METER_READING, A.METER_READING_DTE, A.REAR_AXLE_RATIO,
 E.SERVICE_CENTER_KEY, E.POSITION_DESCR, E.FULL_TM_FLG,
 E.HOURLY_SALARY_FLG, E.ORIGINAL_HIRE_DTE, E.FULL_TIME_HIRE_DTE,
 D.ELOG_TRAINED_FLG, D.ELOG_CERTIFIED_FLG,
 D.LINEHAUL_PACOS_TRAINED, D.PND_PACOS_TRAINED
ORDER BY T.DISPATCH_NBR;
```

Sources

Analysis of Variance from Summary Data. (2017, April). Retrieved November 1, 2017, from http://statpages.info/anova1sm.html

Cerasis IT. “What is LTL Shipping and How Did it Come About?” *Transportation Management Company* | *Cerasis*, 6 Jan. 2017, cerasis.com/2013/11/01/ltl-shipping.

Grace-Martin, K. (n.d.). What is a Logit Function and Why Use Logistic Regression? Retrieved October 25, 2017, from http://www.theanalysisfactor.com/what-is-logit-function/

McNeese, B. (2016, February). Are the Skewness and Kurtosis Useful Statistics? Retrieved October 25, 2017, from https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics

Old Dominion Freight Line. (2017) *2016 Financial Report.* Retrieved from http://www.odfl.com/News/FileServlet?name=Q42016PressRelease.pdf&lib=NewsReleases

Pope, S. (2017, September 1). *OD Technology September 2017 Staff Meeting*. Lecture presented at September 2017 Staff Meeting in Corporate Office, Thomasville.

Raja, V. K., Dhanabal, V., & Chakraborty, G. (n.d.). *Improving performance of Memory Based Reasoning model using Weight of Evidence coded categorical variables* (Tech. No. 10961-2016).