**THE RAYMOND CORPORATION:**

**SCHEDULED MAINTENANCE TIME PERIOD ANALYSIS FOR COUNTERBALANCE MODELS**



**Analysis Conducted and Report Written By**

**Yash Milan Shah, Yiting Zhong, Zhenjie Shen**

**ys889, yz2525, zs382**

**Advised by  
Dr. John Bunge  
  
Submitted to**

**Department of Statistics Science for the Partial fulfilment of the Requirements of the course STSCI 5999: Applied Statistics MPS Project  
  
May 8th, 2019**

**Abstract**

Maintenance a critical topic during the lifetime of a truck. It includes scheduled maintenance and parts repairs. It is important to figure out whether a truck needs to participate in a scheduled maintenance plan, and if it does, how often should the optimal maintenance time periods be.

In this project, two datasets that respectively contain information of counterbalance forklift regular scheduled maintenance as well as other parts and services maintenance, are processed and analyzed. Based on whether trucks participate in SM plans, trucks are divided into SM trucks and Non-SM trucks and compared for differences. Next, for trucks participating in SM plans, baseline scheduled maintenance time periods, such as intervals of 30, 60 and 90 days, are identified from the actual time of days between services. The distribution of interval days is identified as mostly close to burr distribution, and then this variable is classified into seven groups. The differences of variables concerning customer characteristics and truck models are further compared for each interval group.

Besides, the total cost ownership for trucks are found out and compared for different truck demographics and interval groups. K-S test is applied to compare frequency of failure types with/without scheduled maintenance, and it is found that failure types of ELE, LIF and TVL occur less frequently when using a schedule maintenance plan.

Finally, the optimal service time period that leads to minimum costs for each truck are predicted based on truck demographics. Radom forest regression model is applied to do prediction, and after parameter tuning, the model provides a 60% reduction in cost per hour. In order to generate more realistic results, interval groups 1 and 7 are removed in the improved model, and the new model provides a 54% reduction in cost per hour, with 60% trucks assigned to normal interval groups 2 and 3.

Keywords: Truck, Scheduled Maintenance, SM Interval, Cost Ownership

**Contents**

[**Introduction** 4](#_Toc8153378)

[**Glossary:** 6](#_Toc8153379)

[**Methodology** 8](#_Toc8153380)

[**Data Cleaning** 8](#_Toc8153381)

[**1. Scheduled Maintenance Dataset** 8](#_Toc8153382)

[**2. Overall Dataset** 9](#_Toc8153383)

[**Analysis** 11](#_Toc8153384)

[**1. Average SM Interval** 11](#_Toc8153385)

[**Truck Demographics**. 15](#_Toc8153386)

[**Breakdown into SM Groups** 23](#_Toc8153387)

[**Classification of SM Interval groups** 27](#_Toc8153388)

[**2. Cost Ownership Analysis** 28](#_Toc8153389)

[**Meter Hour of Interval Groups** 28](#_Toc8153390)

[**Analysis of Cost (vs Influential Factors)** 29](#_Toc8153391)

[**Cost Ownership Analysis** 29](#_Toc8153392)

[**Analysis of Cost (vs Failure Types)** 33](#_Toc8153393)

[**Optimal Service Time Interval Prediction** 38](#_Toc8153394)

[**Dataset Creation** 38](#_Toc8153395)

[**Model creation**. 40](#_Toc8153396)

[**Conclusions and Recommendations** 42](#_Toc8153397)

[**Conclusion** 42](#_Toc8153398)

[**Optimal Service Time Interval Prediction** 44](#_Toc8153399)

[**Suggestions**. 44](#_Toc8153400)

[**Acknowledgements** 44](#_Toc8153401)

**Introduction**

**About the Company:** TheRaymond Corporation provides end-to-end solutions to help companies better manage their warehousing and distribution, with their headquarters being in Greene, New York. Raymond falls under the Material Handling wing of Toyota. It offers a complete, integrated warehousing and distribution solution to optimize uptime and keep the supply chain going. Industries and Applications of their services are E-commerce and Fulfilment, Food processing, Furniture and Retail, Third Party distribution, Grocery and storage, Medical and Pharmaceutical, and US Federal Government.

**About the trucks and maintenance:** A Counterbalanced Truck is a durable, powerful and highly maneuverable truck. Raymond Counterbalanced Trucks feature a compact footprint and agile steering, making them ideal for confined spaces and close-quarter operations. They are hugely multi-purpose vehicles which are ideal for working alone or with other systems as a part of total material handling solution.  
A scheduled maintenance is a periodic service on a forklift that is provided on a scheduled time period (every 60, 80 or 100 days). The maintenance or service consists of a check lists of brakes, oil steering, etc., much like that of when a car/truck/SUV receives a routine service. Ideal Applications of Counterbalanced trucks are Dock operations, Transport, Put away and Racking.

**Objective:**The objective of this project is to identify current scheduled maintenance plan time periods for Raymond Counterbalance models (overall time between services) and analyze the costs associated with the forklifts.

Once baseline time periods are discovered, attribution of trucks to the appropriate Service time interval groups. Finding the total cost ownership for the Counterbalance Trucks, analyze and compare the total costs between interval groups.

Finally building a model to predict the optimal service time period based on the various factors provided.

**Information the datasets:**The data in this analysis was collected from The Raymond Corporation. The data includes details on Parts and Labor Services for Counterbalanced trucks with various overhead reach types like Stand-Up, Sit-Down, Swing-Reach, Reach-Fork, Side-loaders and Order-pickers.

Based on the2 datasets – Scheduled Maintenance Data, Parts and Labor data, the main goal was to find and analyze scheduled maintenance time periods for the counterbalanced models as above.

The Parts and Labor dataset consisted of data for all Raymond Counterbalanced trucks which can be divided into 2 different sections, ones which have parts and labor repair services, second having parts and labor repair services as well as those which have enrolled for their Scheduled Maintenance plans.

The Scheduled Maintenance dataset consisted of trucks which have parts and labor services and have enrolled for Raymond’s Scheduled Maintenance plans, as well as trucks which have only enrolled for their Scheduled Maintenance plans and have no parts and labor services.  
For cost and time analysis in the second part, the datasets are combined. The table below shows the 3 types of trucks present in the 2 datasets.

Summarizing above there are 3 types of trucks:

|  |  |  |
| --- | --- | --- |
| Type | Parts and Labor Services | Scheduled Maintenance Services |
| 1 | Yes | No |
| 2 | Yes | Yes |
| 3 | No | Yes |

**Glossary:**

The following terms will be used interchangeably in the report: **SM –** Scheduled Maintenance.  
 **Parts** – parts which need to be repaired by Raymond.  
 **SM data –** referral for Scheduled Maintenance data.  
 **Parts & Labor data –** referral for Parts and Labor data repairs.  
 **Raymond Parts repair -** Repairing and replacement of failed parts by Raymond in their factory. These are replaced by parts sold by Raymond.  
 **Labor repair –** Repairing and replacement of failed parts by buying part needed to be replaced from a local shop/external dealer done by Raymond. These are non-Raymond parts which need to be repaired and are bought externally.  
 **esr id –** identification codefor12 different types of Reaches for each truck, out of which 6 are main types mentioned below   
 **Overhead Reach –** The strict structure for each truck. There are 12 types of esr id’s in this data from which finally 6 main types were extracted – SUCB, SIT-DOWN, STAND-UP, 3-WH and PACER. Within each Overhead Reach type, there are different subcategories which are known as esr id’s of the truck. For example, PACER\_<210 and PACER\_210+ are the two types within PACER and only PACER is retained for simplification.  
 **Mast Type –** The type of upright support for each truck. There are more than 100 types of Mast for trucks but they are reduced down to and classified into 5 types by their codes – C30, C40, C50, C60 and Others. For example, C30TT, C30FT are provided in the original dataset, but only C30 is retained as a factor which refers to C30 Mast type.

**Top Customer** **–** Coded as 0/1 dummy variable representing whether a certain customer has the top 10 largest number of total services (SM services as well as Parts Services). Also referred to as Customer Scale.  
 **Failure Type** – There are 141 categories of failure types, but they are classified into 11 main categories after removing number codes and just retaining the type of failure. For example, ELE01, ELE02 are provided in the original dataset, but only ELE is retained as a factor which refers to Electric failure. List of final 11 failure types:

* ELE – Electric
* MEC – Mechanic
* HYD – Hydraulic
* EQU – Equipment
* MOT – Motor
* SM – Scheduled Maintenance
* PUM – Pump
* TVL – Travel
* LIF – Lift-rite
* BAT – Battery
* Others – Other failure types which have less frequency and hence combined into Others.

**SM Interval Groups –** The Average days between Scheduled Maintenance services for trucks were classified into Interval Groups. Detailed definition provided in the next section.  
 **Installation Year –** Installation Year of truck  
 **Service Claim Number –** each service has a unique service claim number which is a code kept for service records.  
 **Line Type –** 1 for Raymond Parts repairs services, 3 for labor repair services, 99 for Scheduled Maintenance services.  
 **Meter Hour –** Meter hour for each truck to date from the time of its manufacture date.  
 **Max Meter Hour** – Meter hour on the last date of service/max meter hour recorded in a truck’s lifetime. (These values are used in the summarized datasets below)  
 **Date** – Date in dd/mm/yyyy format and time in HH:MM:SS format for each service. This is the recorded date of the work order/service took place.  
 **Quantity –** Quantity of parts or labor used for each repair service. No quantity of parts used mentioned for trucks in the SM dataset.  
 **Amount –** Current price of parts repaired by Raymond. Amount mentioned for Raymond parts repair (refer above), and NA for Labor repairs (refer above). No repair amount mentioned for trucks in the SM dataset.  
 **Cost per hour –** Cost per meter hour of a truck.

**Methodology**

**Data Cleaning:**

**1. Scheduled Maintenance Dataset:** The dataset consists of more than 120,000 observations having only SM services, with multiple entries for each Truck ID, each representing an observation/service in detail with 30 features including the truck demographics like Truck ID, Installation Year, Model Number, Mast Type, esr id, Overhead Reach type and there other service demographics like Service Claim Number, Transaction type, Line Type, Date, Quantity and Amount.

The Quantity and Amount for this dataset are provided with blank/null values. For the cleaning of this dataset, the first step involved changing wrong values of some customer states as their state codes were incorrect. For example, states were named “XX” instead of their respective state codes and these were changed by logical imputation based on their customer cities or zip codes.

The following step involved sorting the trucks by truck ID’s and then by dates within each truck ID. This dataset only represents the Scheduled Maintenance service observations for each truck and does not contain the first date of any part repair for each truck. For each truck ID, the first date part repair information for each truck ID present in the SM dataset was looked up from the Parts & Labor dataset to get a baseline date (date of first ever repair to occur) for each truck after which the Scheduled Maintenances can take place.

Following the above step, after sorting the dataset by truck ID and date within each truck ID, the difference in days between each service was calculated for each truck. Some trucks had a difference of 5 to 6 years between 2 services due to change in client or due to irregular SM services and hence had outlier numbers of higher than 1000 days between 2 services.

The next step involved cleaning the data to extract Overhead Reach types from esr id’s for each truck ID, and extracting Mast Types as mentioned above. These were one using the gsub function along with the “stringi” package in R. Extra characters in the Overhead Reach type, and Mast Type were removed. Accurate Mast Type’s were extracted from truck’s unique 16 digit model number consisting of Overhead Reach + Mast Type + Model Number + Truck ID.

This makes the dataset ready to be summarized into a smaller aggregated dataset having just one observation per Truck. This was done by calculating the Average time between services for each truck ID and summarizing it into a smaller dataset having 7086 observations (of which all have only SM services or SM services along with parts & labor services).

**2. Overall Dataset:** The Parts and Labor dataset consists of about 700,000 observations having trucks with Parts and Labor Services only and trucks with Parts and Labor Services along with SM services. The dataset has the same features as the SM dataset for each truck.

This dataset was then combined with the Scheduled Maintenance dataset, together having around 900,000 observations.

First step was again changing wrong values of some customer states as their state codes were incorrect. For example, states were named “XX” instead of their respective state codes and these were changed by logical imputation based on their customer cities or zip codes.

The Amounts for Raymond parts services had null values for many of the observation. This was changed by mean imputation for each failure type. For example, the amount for trucks having failure type Electric (ELE) has a mean amount of $114 for all such trucks, excluding the null values, and hence the null values were replaced by this mean value. This was done for all the trucks having Raymond Parts services. The Amounts for Labor part services had null values for all services, and these were replaced by a standard value of $100 for each truck with labor part services. Finally, cost was found for each truck, which was a product of Amount and Quantity of parts sold.

The meter hours for each truck were provided with many unrealistic/inconsistent values which were changed to median values. The 3rd quartile for meter hour values was around 7000 hours and after which in the last quartile of the data, there were values ranging from 20,000 to 1,000,000. All unrealistic values above 30,000 meter hours were changed to the median value of meter hours as per discussion with the Raymond team.

The duplicate entries for a particular day were removed, i.e. the data was aggregated down to 250,000 observations by clubbing multiple service entries for each truck on the same day at the same time. Hence, the cost was aggregated for all such services on the same day at the same time. The problem on hand was that some services have same service numbers but are a follow-up of the previous service and hence their time is different on the same day, and some services do have same date and time, but they are for repairing different failed parts. This was overcome by forming a unique identifier by merging the service claim number, date and time of the service. Failure types were made into separate features and a total count of a failure type for each truck was counted, to avoid loss of information while removing the duplicate service entries on the same date.   
  
For example, the first table was reduced to the second table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Truck ID** | **Service Claim** | **Date** | **Part failed** | **Line type** | **Cost** |
| 12345 | D12345 | 12-09-2017 | ELE | 1 | 213 |
| 12345 | D12345 | 12-09-2017 | MEC | 3 | 100 |
| 12345 | D12345 | 12-09-2017 | MEC | 1 | 314 |
| 9876 | D9876 | 03-01-2018 | MOT | 1 | 455 |
| 9876 | D9876 | 03-01-2018 | MOT | 1 | 42 |
| 456456 | D456456 | 04-05-2018 | HYD | 3 | 100 |
| 456456 | D456456 | 04-05-2018 | PUM | 1 | 4500 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Truck ID** | **Service Claim** | **Date** | **Line type** | **Cost** | **ELE** | **MEC** | **MOT** |
| 12345 | D12345 | 12-09-2017 | 1 | 213 | 1 | 2 | 0 |
| 9876 | D9876 | 03-01-2018 | 3 | 100 | 0 | 0 | 2 |
| 456456 | D456456 | 04-05-2018 | 1 | 314 | 0 | 1 | 1 |

The following step involved sorting the trucks by truck ID’s and then by dates within each truck ID. This involved all the services after clubbing the 2 datasets. After sorting the dataset by truck ID and date within each truck ID, the difference in days between each service was calculated for each truck. Some trucks had a difference of 5 to 6 years between 2 services due to change in client or due to irregular services and hence had outlier numbers of higher than 1000 days between 2 services.

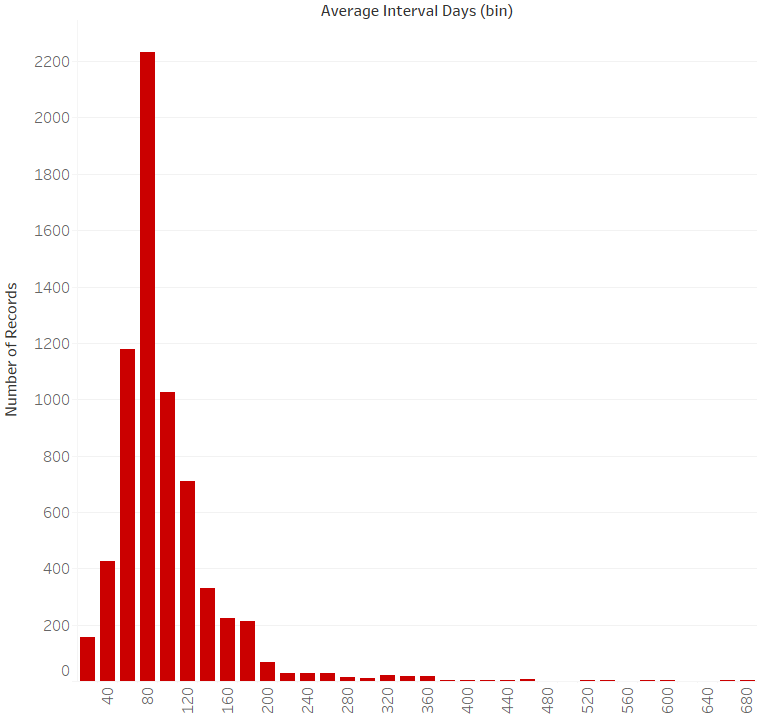
The trucks were now classified into 3 categories which were mentioned above: Both, SM and Non-SM.

After finding average days between services for this dataset, it was summarized down a smaller aggregated dataset having just one observation per Truck. This was done by calculating the Average time between services for each truck ID and summarizing it into a smaller dataset having 8320 observations (of which 7014 are trucks with both services, 72 with only SM services and 1234 trucks with No SM services).

**Analysis**

**1. Average SM Interval (Only SM Trucks):** After summarizing the dataset to 7086 observations and 18 features including truck demographics as mentioned earlier and service details like parts failed, cost, top states in which services are provided, top customers, etc.

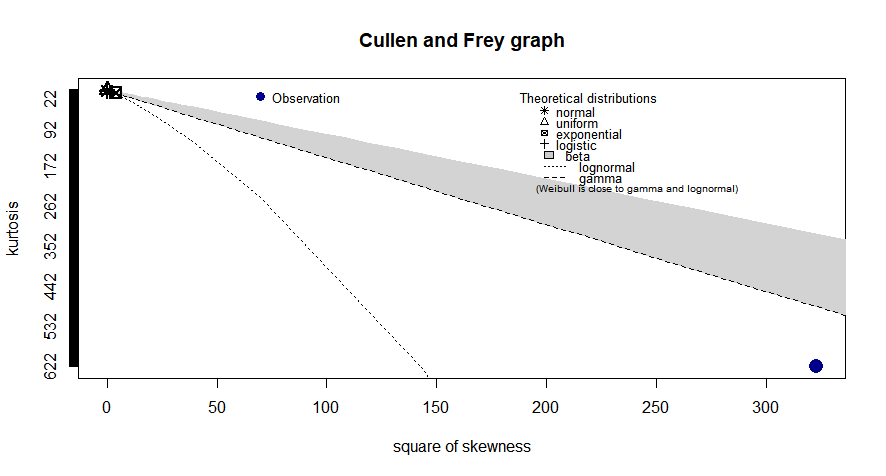
Maximum meter hour for each truck was noted and Cost per hour (meter hour) and services per hour were calculated for each truck.

In order to find the average SM interval, the dataset was sorted by service time of the trucks in chronological order, and the gaps between each two consecutive services were calculated. In addition, on some days, there could be multiple service times for the same truck, hence resulting a 0-day interval. In this analysis, if more than one service times exist for the same truck within a 24-hour time period, then only one service time will be used in calculating the average SM interval days for that particular truck.   
Figure 1 is the distribution of Average SM Interval days for each truck:  
  
****  
 Figure 1Average Intervals for each truck are the average number of days between all services for that particular truck.

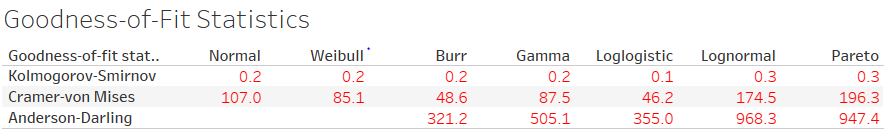
The plot has a very elongated right tail due to the presence of outliers as there is a lot of gap in time between Scheduled Maintenances, i.e. higher average intervals between services for trucks.

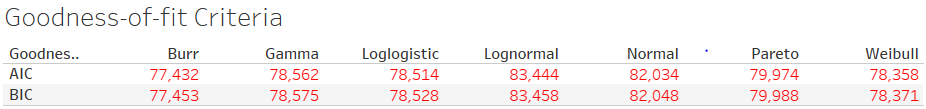
One reason is change of customer name when a new date occurs, and another is having less number of services, i.e. the second service after the first repair occurs after a long time interval. For example, one service may be for client named “ABC” on 08-02-2008 and the next service would be for a client named “BCF” on 09-12-2016, due to which the average number of days between services for that truck would be influenced by this gap between 2 services.

After running a few statistical tests below, it’s confirmed that the best fitting distribution for the Average Interval data (Time in days between services) is the Burr distribution.  
The 3 plots below show a comparison between the Burr and Gamma distributions as these are the 2 best fitting distributions to the Average SM Interval between services.

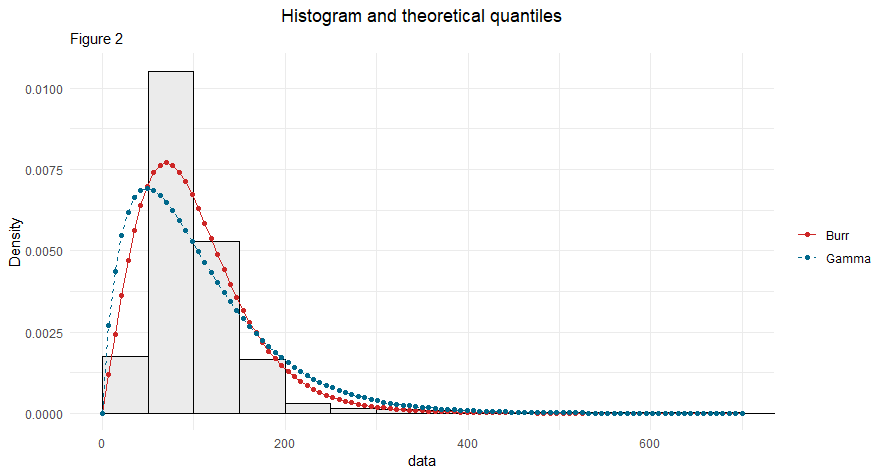
The Cullen and Frey graph shows that the distribution lies somewhere near the Gamma and Weibull distributions, while other tests confirm that it is closer to Burr distribution than Gamma or Weibull. Following is a plot for the Cullen and Frey graph: 

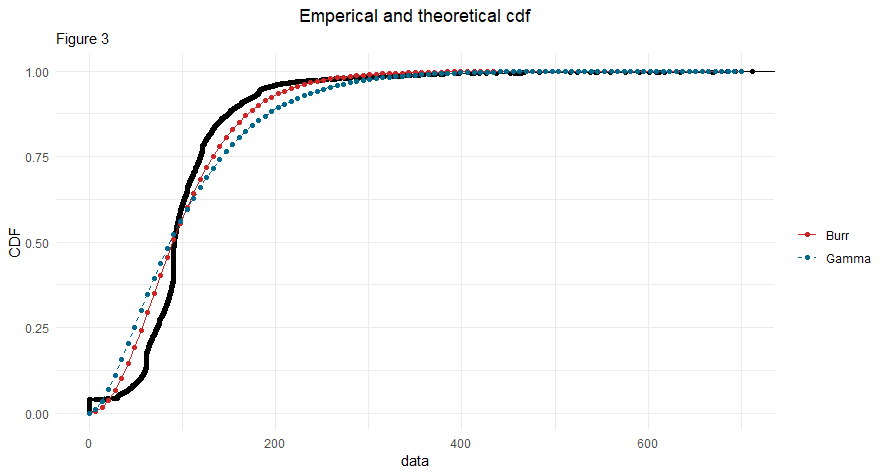
Below are tables for goodness-of-fit statistics and goodness-of-fit criteria tests conducted on the Average SM Interval data:

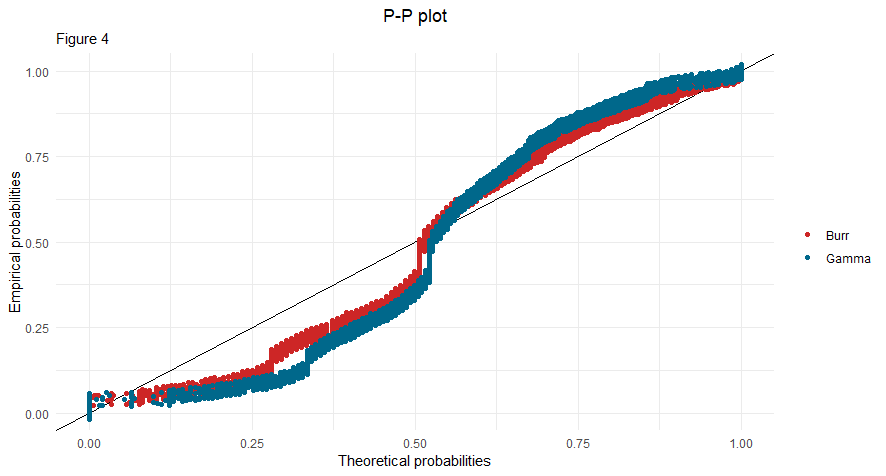




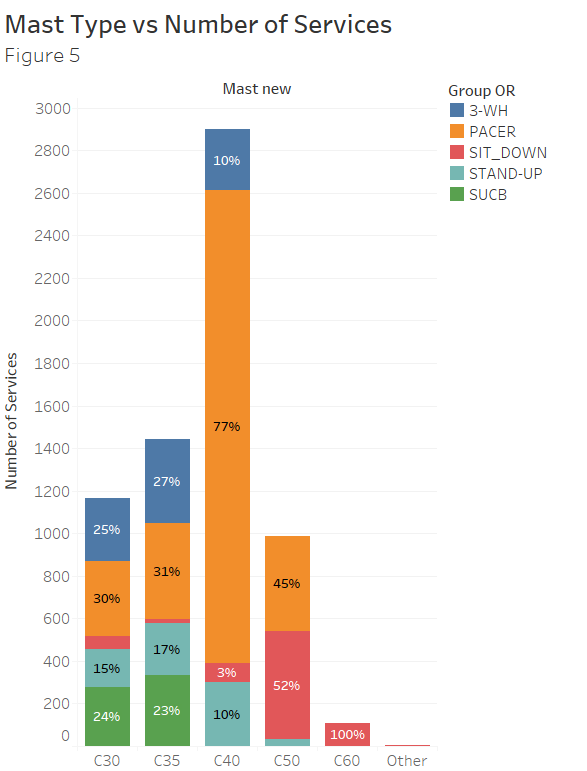
The AIC and BIC show that Burr distribution is a better fit for this data compared to the other distributions.

The Gamma and Weibull distributions are close to the Burr distribution as well.  
Below are some plots for comparing the Burr and the Gamma distributions:  
  


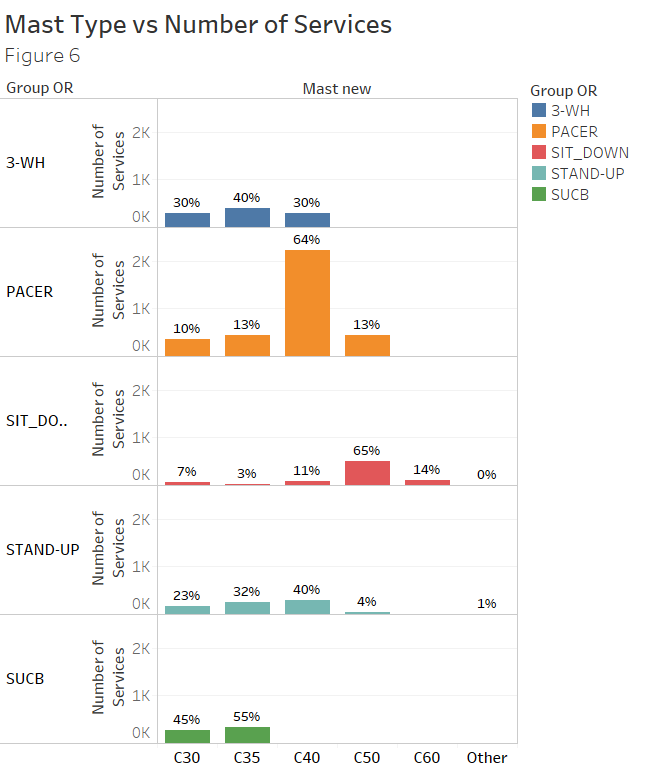


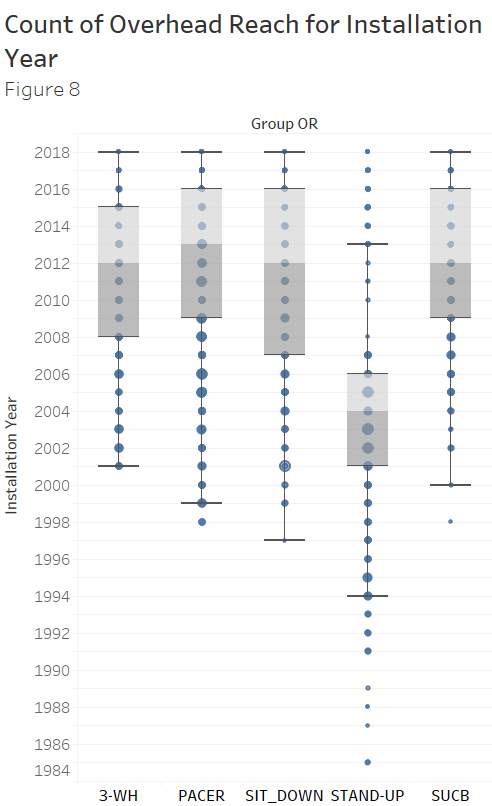


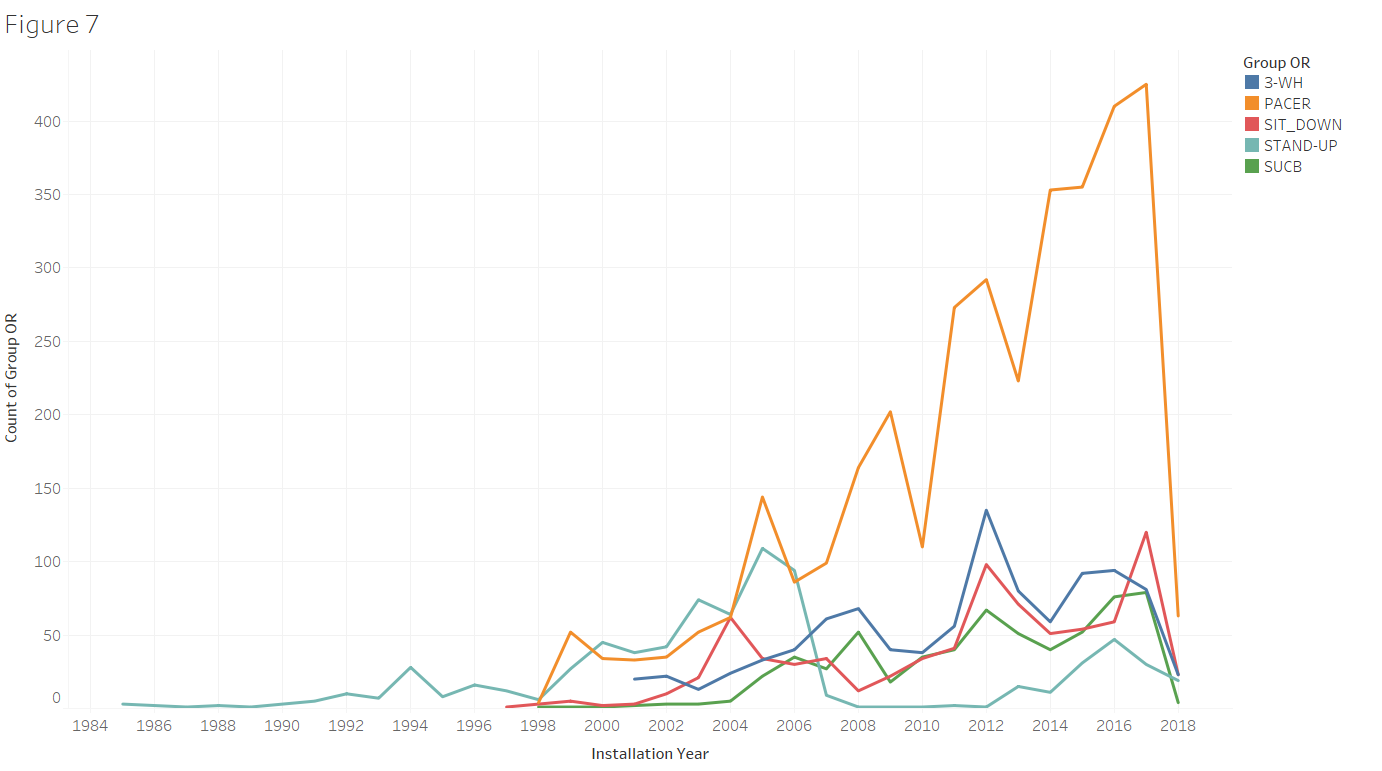
**Truck Demographics:** 1. Comparing the number of services amongst each Mast type, and how they are distributed within each Overhead Reach type, within these different Mast types. Figure 5 shows the Mast Types distribution against number of services and how they’re distributed for each of these different overhead reach types.



It is visible from Figure 5 that the trucks with C40 Mast types are the most frequently serviced ones. It’s also visible that the PACER type is the most common within all the different Mast types, except for the C50 type where SIT-DOWN types are the most common ones. The C60 type only consist of SIT SOWN type Overhead reaches.

2. Comparing the number of services amongst each Overhead Reach type, and how they are distributed within each Mast type for each of these different Overhead reach types. Below is a plot, Figure 6 which shows within each overhead reach type, how the Mast types are distributed.  
  
  
The C40 mast type is the most commonly serviced one within the PACER and STAND-UP type trucks, whereas C35 is the most commonly serviced one within 3-WH and SUCB type trucks, and C50 services dominates the SIT-DOWN type trucks.





3. Figures 7 and 8 are a comparison for the Count of Overhead Reach types versus their Installation Years. The STAND-UP trucks were introduced in the 1980s whereas others only originated after late 1990s or 2000s.

The thickness of the circles (points) in Figure 7 for each truck depends on the number of services during that particular year. Thick circles mean there were more services for that Overhead Reach type during that particular year.

4. Following this, there was an analysis conducted on whether all forklifts participate in the Scheduled Maintenance programs, and out of the ones that do, if all have regular services.

Some basic statistics comparing trucks having SM services with trucks having only Parts and Repairs services (Not SM Trucks).

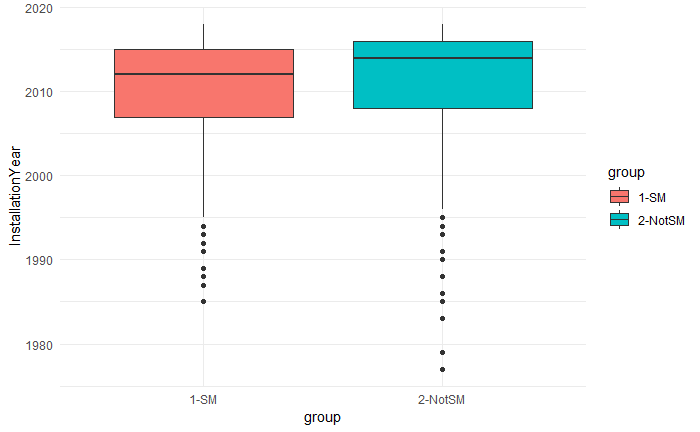
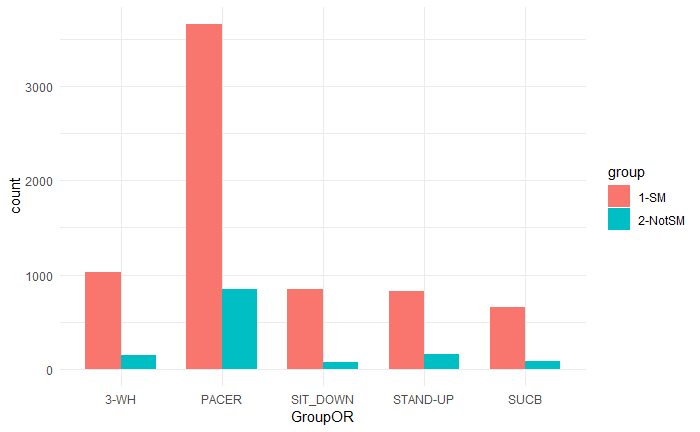
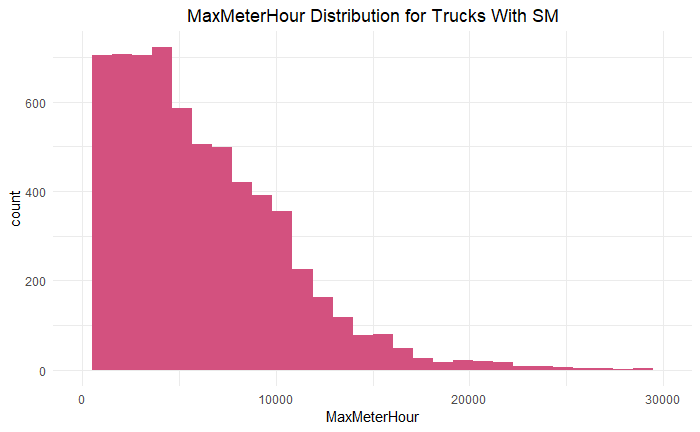
  
 Figure 9

Figure 9 shows that the Not SM trucks are slightly newer than the SM trucks.

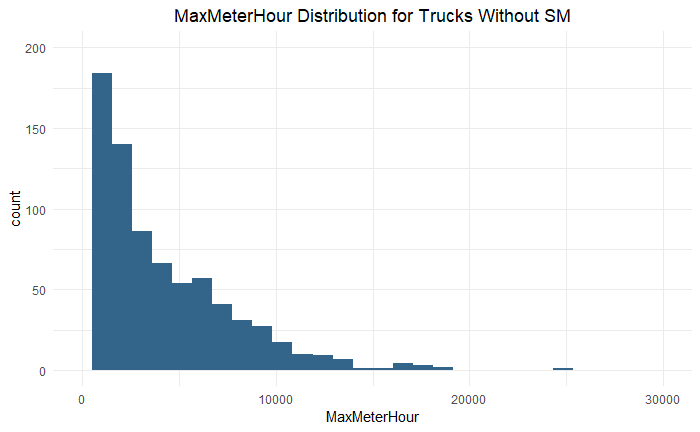
  
Figure 10

As mentioned earlier, Figure 10 agrees with the conclusion that SM trucks are much more in frequency (7086) than Not SM trucks (1234). For Overhead Reach, PACER is the most common, 3-WH and STAND-UP have a higher proportion in Not SM trucks.

Below are plots comparing the Max Meter Hours for trucks SM trucks and Not SM trucks.

  
 Figure 11

Maximum Meter Hour for SM trucks are more concentrated in (0, 10000), especially in (0,5000) hours range

  
 Figure 12

Non-SM trucks have generally fewer meter hours, with most of them concentrated in (0,2500) hours range

**SM Distribution by State:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **State** | | **Service Count** | **SM Rate** | **State** | **Service Count** | **SM Rate** | **State** | **Service Count** | **SM Rate** | **State** | **Service Count** | **SM Rate** |
| GA | 206951 | | 98.1% | MD | 84 | 100.0% | PA | 18 | 100.0% | NJ | 3 | 66.7% |
| SC | 184266 | | 98.8% | WI | 83 | 100.0% | AR | 18 | 100.0% | UT | 3 | 0.0% |
| NC | 176562 | | 98.3% | TX | 74 | 100.0% | OK | 9 | 0.0% | IN | 3 | 100.0% |
| AL | 50091 | | 96.9% | TN | 66 | 69.7% | KY | 6 | 100.0% | WA | 3 | 100.0% |
| FL | 4573 | | 99.1% | OH | 57 | 91.2% | NV | 5 | 80.0% | MO | 2 | 0.0% |
| CA | 242 | | 89.3% | MN | 50 | 88.0% | NE | 5 | 100.0% | CO | 1 | 100.0% |
| IA | 220 | | 100% | NY | 36 | 97.2% | MI | 5 | 100.0% |  |  |  |
| IL | 110 | | 70.0% | AZ | 24 | 0.0% | MA | 5 | 100.0% |  |  |  |

The above table shows the SM rate for services provided in each state. It is arranged by descending order of Service Count.

Georgia, South Carolina and North Carolina have the highest services counts, with SM rates being higher than 98%.

The Top 10 states have a high SM rate.

There are some states with very few services which do not participate in the SM plans.

**SM Interval Influential Factor Analysis – Regression Model:** A Linear regression model was built in order to explore the important influential factors affecting the SM Interval. In the model, the response variable is Average SM Interval Days, and the predictors include Installation Year, Customer Scale, Customer State, Mast type and Overhead Reach.

The model results are shown in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Estimate** | **Standard Error** | **P Value** | **Significance** |
| Installation Year | -2.434 | 0.221 | 0.000 | \*\*\* |
| Top Customer | -7.961 | 4.099 | 0.052 | . |
| State\_CA | -2.610 | 22.391 | 0.907 |  |
| State\_FL | 51.714 | 9.988 | 0.000 | \*\*\* |
| State\_GA | 14.306 | 3.644 | 0.000 | \*\*\* |
| State\_IA | 1376.245 | 44.337 | 0.000 | \*\*\* |
| State\_IL | 14.947 | 88.453 | 0.866 |  |
| State\_MD | 234.026 | 62.563 | 0.000 | \*\*\* |
| State\_NC | -68.361 | 88.398 | 0.439 |  |
| State\_NJ | 4.099 | 3.800 | 0.281 |  |
| State\_NY | -8.309 | 24.757 | 0.737 |  |
| State\_PA | 55.823 | 88.468 | 0.528 |  |
| State\_SC | 2.572 | 3.833 | 0.502 |  |
| State\_TX | 11.434 | 88.363 | 0.897 |  |
| Mast Type\_C35 | 39.130 | 88.414 | 0.658 |  |
| Mast Type\_C40 | -14.220 | 3.374 | 0.000 | \*\*\* |
| Mast Type\_C50 | -5.975 | 3.265 | 0.067 | . |
| Mast Type\_C60 | -13.175 | 4.470 | 0.003 | \*\* |
| Mast Type\_Other | 9.305 | 9.111 | 0.307 |  |
| GorupOR\_PACER | -9.700 | 3.387 | 0.004 | \*\* |
| GorupOR\_SITDOWN | -1.235 | 5.168 | 0.811 |  |
| GorupOR\_STANDUP | -19.362 | 4.404 | 0.000 | \*\*\* |
| GorupOR\_SUCB | 10.153 | 4.507 | 0.024 | \* |

Several variables show statistical significance under 0.05 level in this model and the interpretations of parameter estimates are as below:

* When installation year of trucks increases by 1 year, their average SM interval decreases by 2.4 days.
* Top 10 customers generally have their trucks scheduled maintained 8 days more frequently than other customers.
* The baseline group for Customer State is Albama (AL). Compared to AL, states Florida (FL), Georgia (GA) and Iowa (IA) do not participate in SM services very often.
* The baseline group for Mast Type is C30. Compared to C30, Mast Type C40, C50, C60 are more frequent in SM services.
* The baseline group for Overhead Reach is 3-WH. Compared to 3-WH, Overhead Reach PACER and STAND-UP are more frequent in SM services and SUCB is less frequent in SM services.

**Breakdown into SM Groups:** These Interval Groups are assigned to all trucks depending on their Average Interval Days between Scheduled Maintenance Services.

By checking the distribution (quantiles, median) of average SM interval, the Average Interval Days were split into 7 groups with these intervals where Group 1 represents trucks with no SMs as of year to date. The groups are as below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Interval (Days)** | **Interval Group** | **Count** | **%** |
| 0 | 1 | 283 | 4% |
| (0-60] | 2 | 651 | 9% |
| (60-90] | 3 | 1898 | 27% |
| (90-120] | 4 | 2512 | 35% |
| (120-180] | 5 | 1248 | 18% |
| (180-240] | 6 | 298 | 4% |
| (240+) | 7 | 196 | 3% |

The SM Intervals are now categories. Below are some ANOVA analyses and chi-squared tests to test for differences between SM groups on Installation Year, Mast Type, etc.

The ANOVA analysis performed on Installation Year has a very small overall P-value. To check which SM Interval groups have a significant difference, pairwise comparisons are performed to compare two specific SM Interval groups based on the Installation Year of trucks in it.

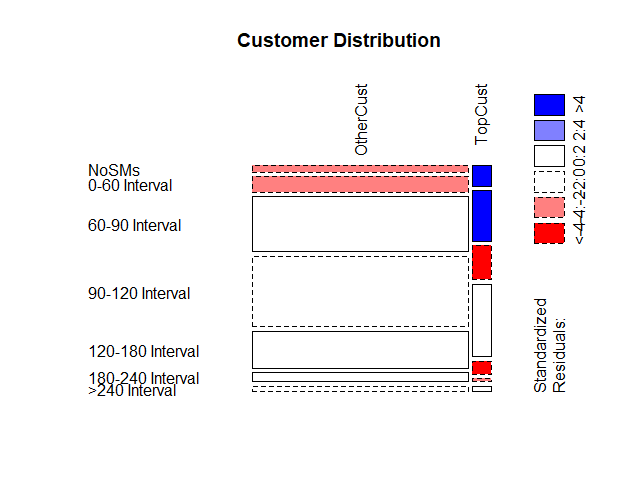
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Group** | **Difference** | **Lower Bound** | **Upper Bound** | **Adjusted P Value** |
| 2 – 1 | -2.544 | -3.659 | -1.430 | 0.000 |
| 3 – 1 | -2.715 | -3.712 | -1.717 | 0.000 |
| 4 – 1 | -2.597 | -3.578 | -1.615 | 0.000 |
| 5 – 1 | -3.467 | -4.498 | -2.436 | 0.000 |
| 6 – 1 | -3.162 | -4.462 | -1.863 | 0.000 |
| 7 – 1 | -5.343 | -6.798 | -3.888 | 0.000 |
| 3 – 2 | -0.170 | -0.881 | 0.541 | 0.992 |
| 4 – 2 | -0.053 | -0.741 | 0.636 | 1.000 |
| 5 – 2 | -0.923 | -1.680 | -0.166 | 0.006 |
| 6 – 2 | -0.618 | -1.713 | 0.477 | 0.640 |
| 7 – 2 | -2.799 | -4.074 | -1.523 | 0.000 |
| 4 - 3 | 0.118 | -0.358 | 0.594 | 0.991 |
| 5 - 3 | -0.752 | -1.323 | -0.182 | 0.002 |
| 6 - 3 | -0.448 | -1.423 | 0.528 | 0.826 |
| 7 - 3 | -2.628 | -3.803 | -1.454 | 0.000 |
| 5 - 4 | -0.870 | -1.412 | -0.328 | 0.000 |
| 6 - 4 | -0.565 | -1.524 | 0.394 | 0.590 |
| 7 - 4 | -2.746 | -3.907 | -1.585 | 0.000 |
| 6 - 5 | 0.305 | -0.705 | 1.314 | 0.974 |
| 7 - 5 | -1.876 | -3.079 | -0.673 | 0.000 |
| 7 - 6 | -2.181 | -3.620 | -0.741 | 0.000 |

As shown above in the table, Small P-values show that Group 1, 5 and 7 are significantly different from the other groups based on Installation Year.

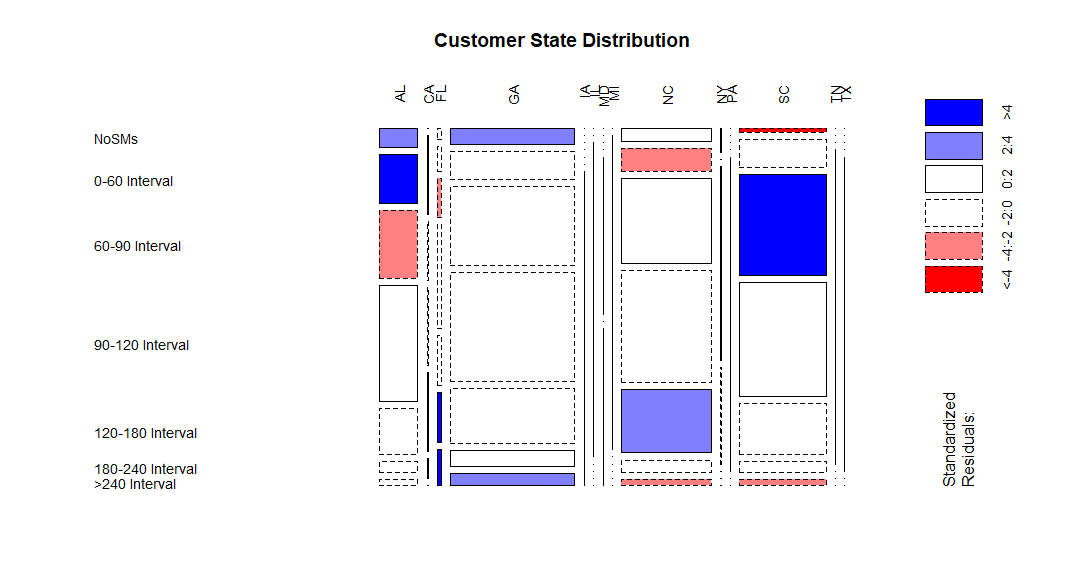
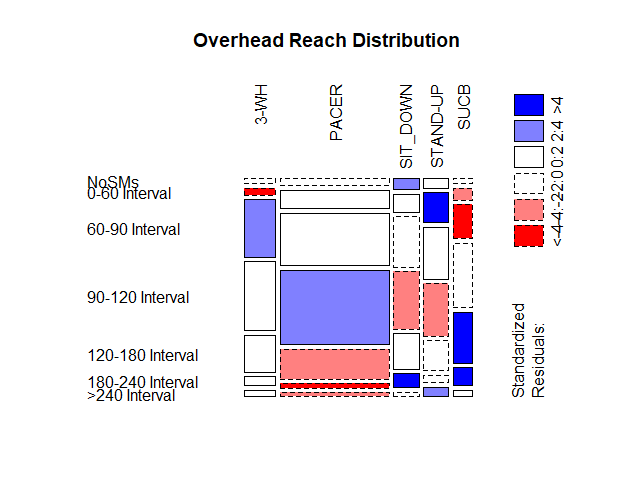
The median Installation Year for Group 1, 5, 7 are 2017, 2011 and 2018 respectively, with other groups having median Installation Year around 2012. Notice that Group 7 has an obvious drop in the Installation Year because a large proportion of outliers exist in this group. As mentioned in the report before, for many trucks installed in 2008, there is significant time gap between two SM services causing these trucks falling into the 7th group.

**Further Analysis:** Mosaic plots and chi-squared tests are performed to test difference in SM Interval Groups for discrete variables like Customer Scale, Customer State, Overhead Reach and Mast Type.

Mosaic Plot clearly recognizes relationships between different categorical variables. It is used to compare the expected frequency versus observed frequency. In these plots, blue color indicates that the observed value is higher than the expected value if the data were random, and red color specifies that the observed value is lower than the expected value if the data were random.

  
 Figure 13

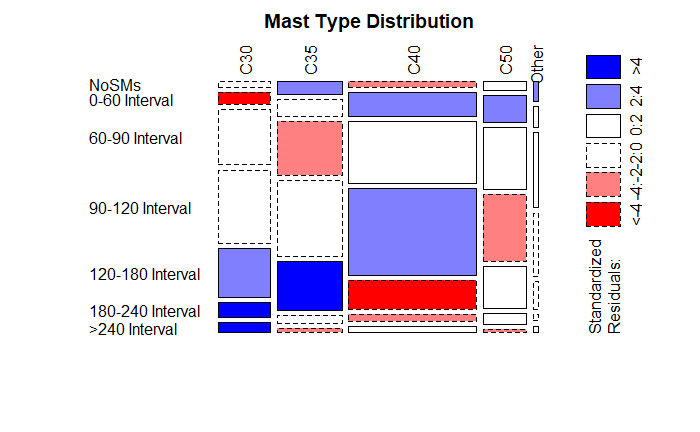
Top 10 customers in services are more likely to have frequent scheduled maintenance plans

  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
 Figure 14  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
 Figure 15

Truck models of STAND-UP and 3-WH are more likely to have frequent schedule maintenance.

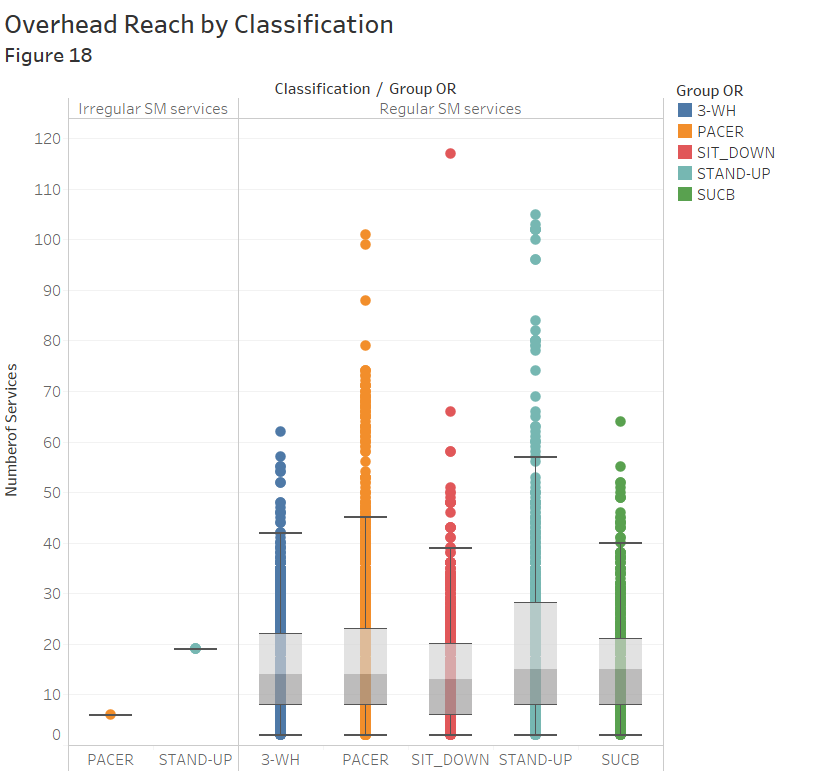
Models of SIT-DOWN and SUCB are less likely to have frequent schedule maintenance services.

Customers in AL, SC are more likely to frequently schedule maintenance, customers in FL, NC are less likely to frequently have scheduled maintenance services.

  
  
  
  
  
  
  
  
  
  
  
  
  
Figure 17

Mast Type of C40 and C50 are more likely to have frequent SM plans.

Mast Type of C30, C35 are less likely to have frequent SM plans.



**Classification of SM Interval groups:** The plot above shows that within the trucks having SM plans, the trucks are classified into regular SM or irregular SM services by their Average Interval between services.

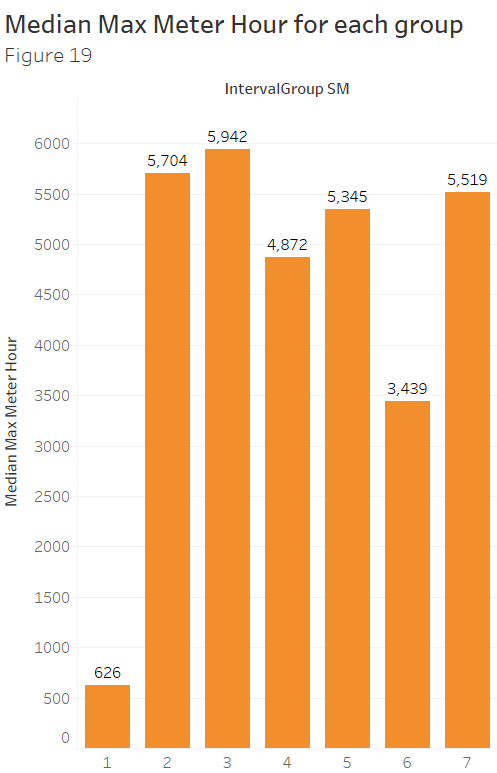
The trucks are divided into 7 interval groups as mentioned and then classified into 2 groups based on regular or irregular services.

One being Regular SM services and one being Irregular SM services (too much gap in days between 2 services).

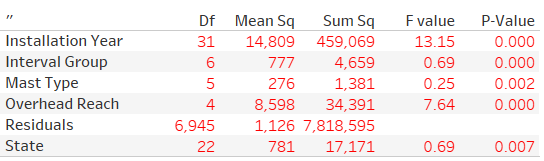
The latter group also consists of outliers having very few and hence irregular services.

**2. Cost Ownership Analysis (Trucks with Only SM and SM and Parts Services):**

**Meter Hour of Interval Groups:** Before moving on to cost ownership (cost per hour) analysis, we first see the Median Meter Hour for all interval groups, Figure 19 below shows a comparison of Median Max Meter Hours of the groups. Group 1 has very low median meter hours for it’s trucks compared to the other groups as they have only 1 service to date. Group 3 has the highest median meter hour compared to the rest but all seem close above 5,300 hours, except for Group 1,4 and 6.



**Analysis of Cost (vs Influential Factors):**



The table above illustrates the summary of an overall ANOVA analysis of the potential factors for cost per hour; each of the five chosen factors appears to be very significant in terms of affecting cost per hour of the trucks. Results of more detailed analysis are discussed in the following sections.

**Cost Ownership Analysis**

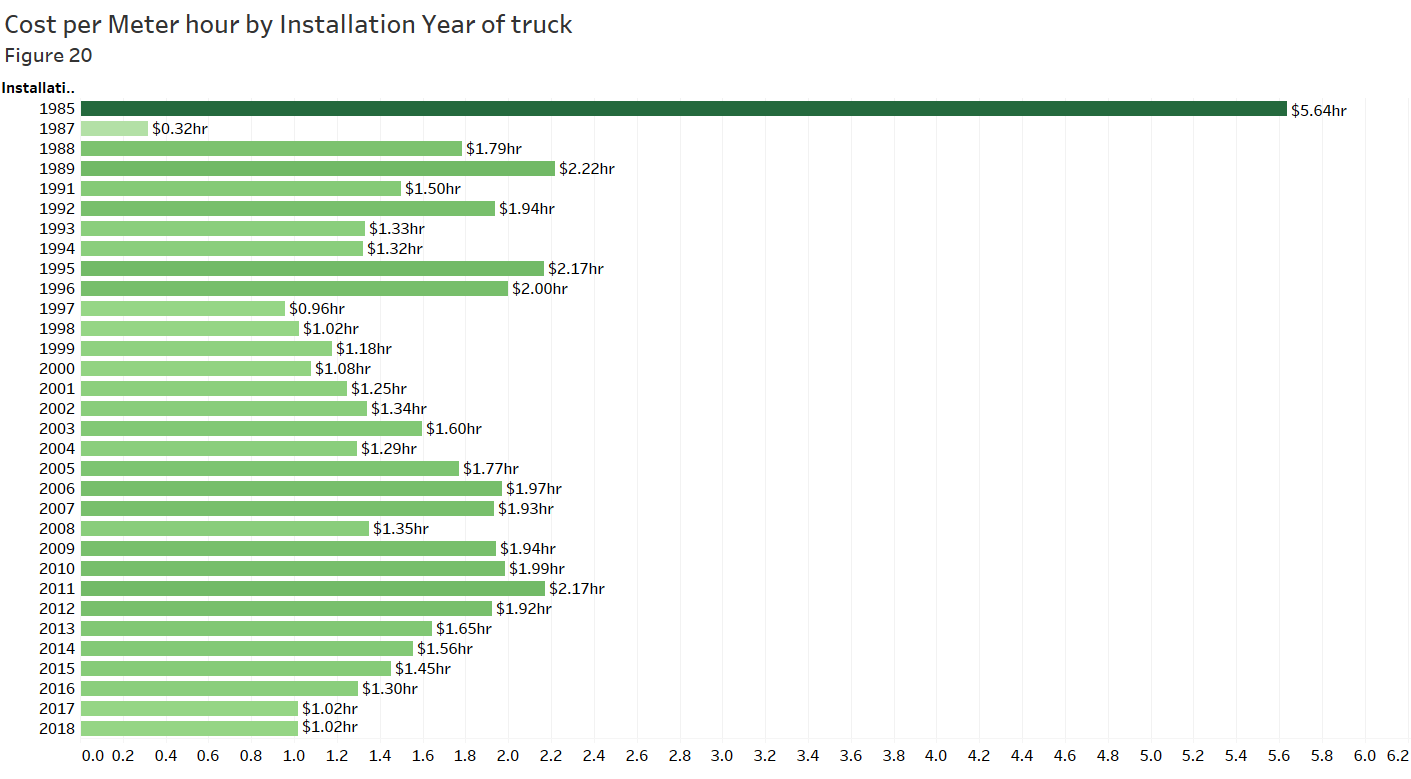


Figure 20 shows the median cost per hour of the trucks against their installation years. Here median (rather than mean) was used since it will not be affected by the existence of the outliers, and when total cost per hour is used, it is highest for years where more trucks were manufactured.

The trucks installed in the year 1985 have the highest median cost per hour (which is affected by outlier entries along with starting meter hours of trucks being less during that time).

Overall, there is a decreasing trend from 1985-2018, namely newer trucks tend to have lower values of cost per hour of operation.

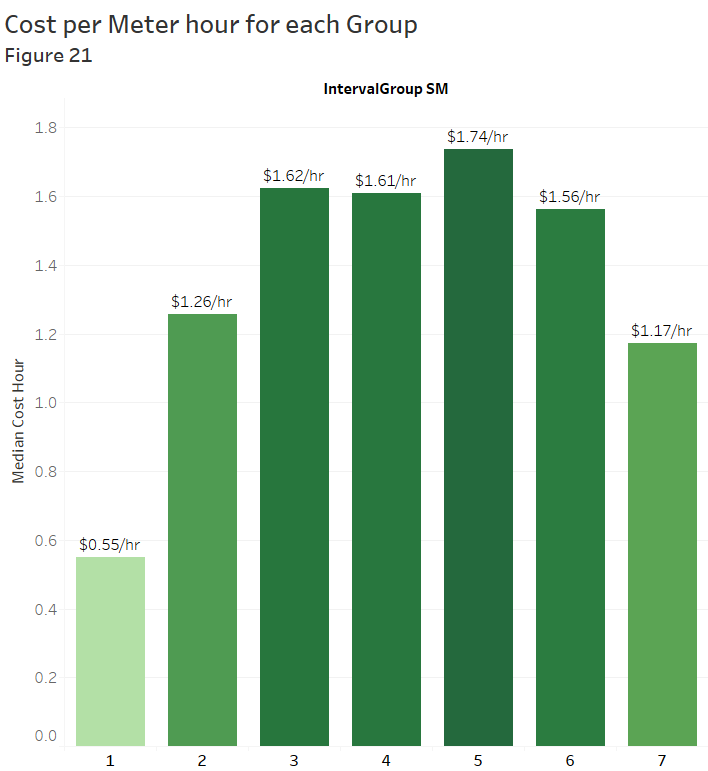
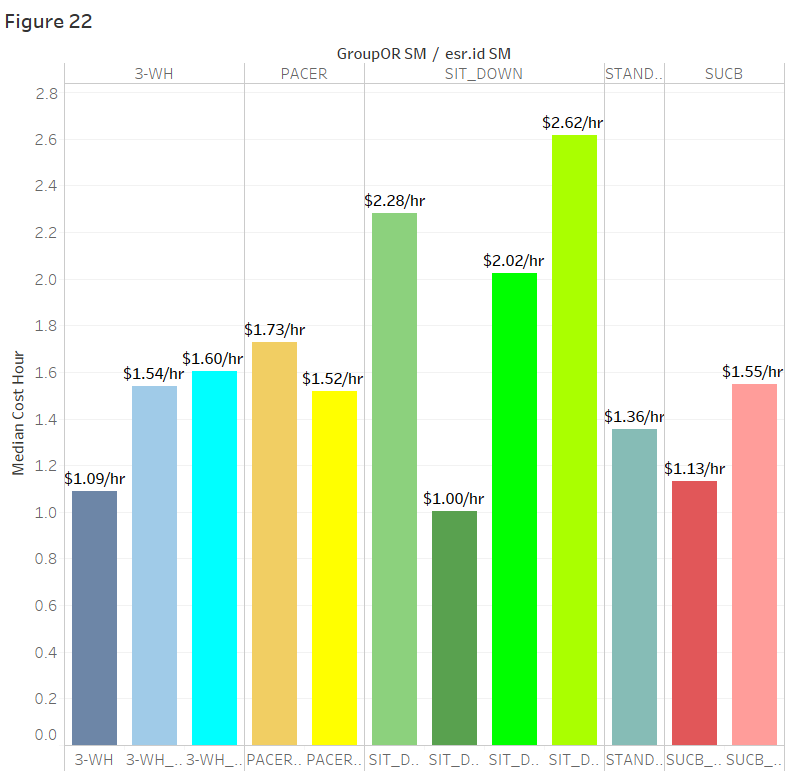
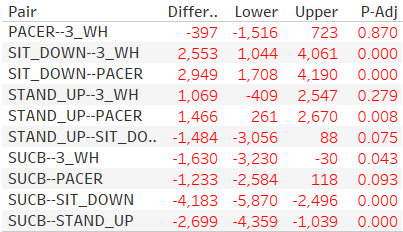


Figure 21 breaks down the trucks into SM Interval Groups and shows the median cost per hour for each group. Group 1 has the lowest median, which makes sense as it consists of the trucks that have 0 or just 1 service. Group 7 is also a group that basically consists of outliers, and it has the second lowest median. Groups 2-6 are those trucks that have regular services, and it is not surprising to see these groups with higher medians.

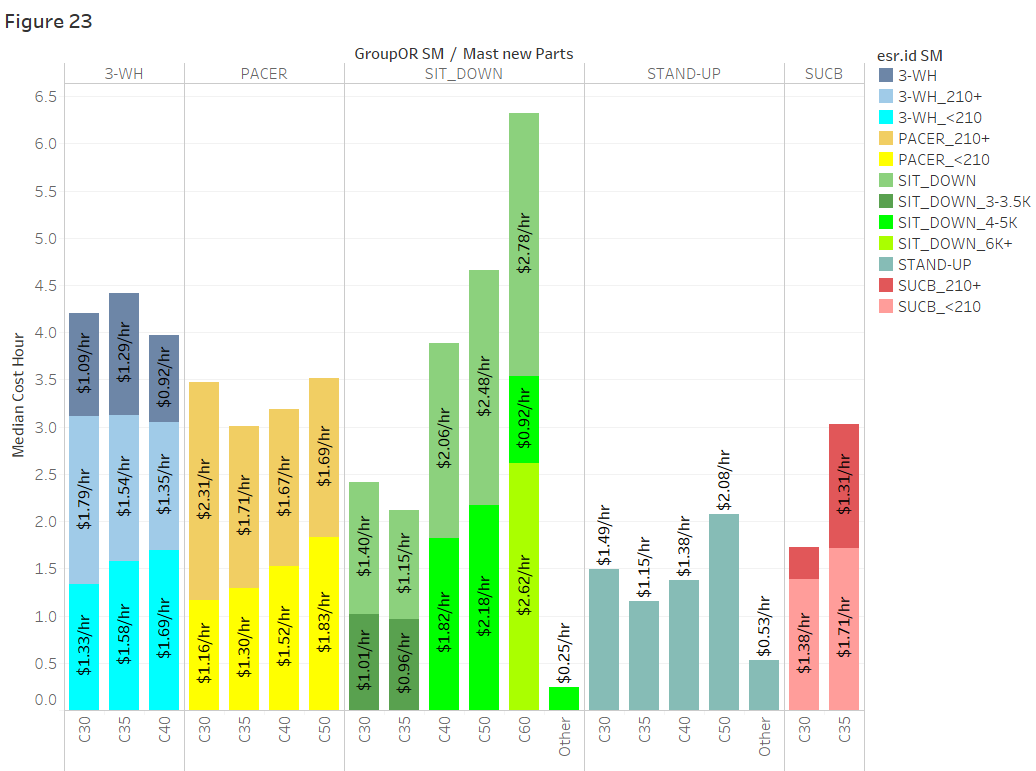


Next, in Figure 22 the trucks were broken down by their Overhead Reach (OR) types, as well as sub-OR types (esr id). A pair-wise comparison table is also provided below.

The highest median cost per hour was for the SIT-DOWN trucks compared to the rest, where the SIT-DOWN\_6K+ truck has the highest median cost per hour which is $2.62 per hour. The SUCB trucks overall had the lowest median cost per hour. But within the 3 most frequent serviced trucks which are PACER, SIT-DOWN and STAND-UP trucks, the SIT\_DOWN trucks have the highest median cost per hour.



By looking through the p-values, it appears that SIT-DOWN type has significantly higher costs than all other types, STAND-UP type has significantly higher costs than types PACER and SUCB, and 3-WH type has significantly higher costs than type SUCB. Note the agreement between the plots and the pairwise comparisons.



In Figure 23, the trucks were further broken down by Mast Type within each main Overhead Reach group, and for each Mast Type bar, the median cost of each sub-Overhead Reach group is shown as a stack of the bar. For instance, within the SIT-DOWN types, Mast Type C60 trucks have the highest median cost per hour, and there are three contributing sub-SIT-DOWN types for this particular median.

**Analysis of Cost (vs Failure Types):**

This section of analysis is aimed to answer the following two questions:

* What types of repairs seem to occur less frequent when using a SM plan?
* What types of repairs seem to occur less frequent by decreasing time interval in an SM plan?

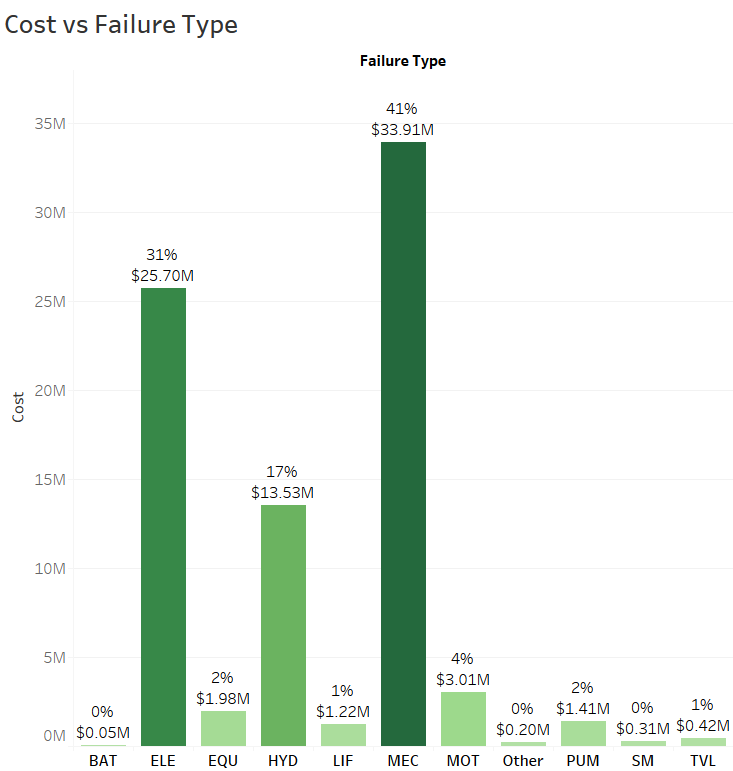
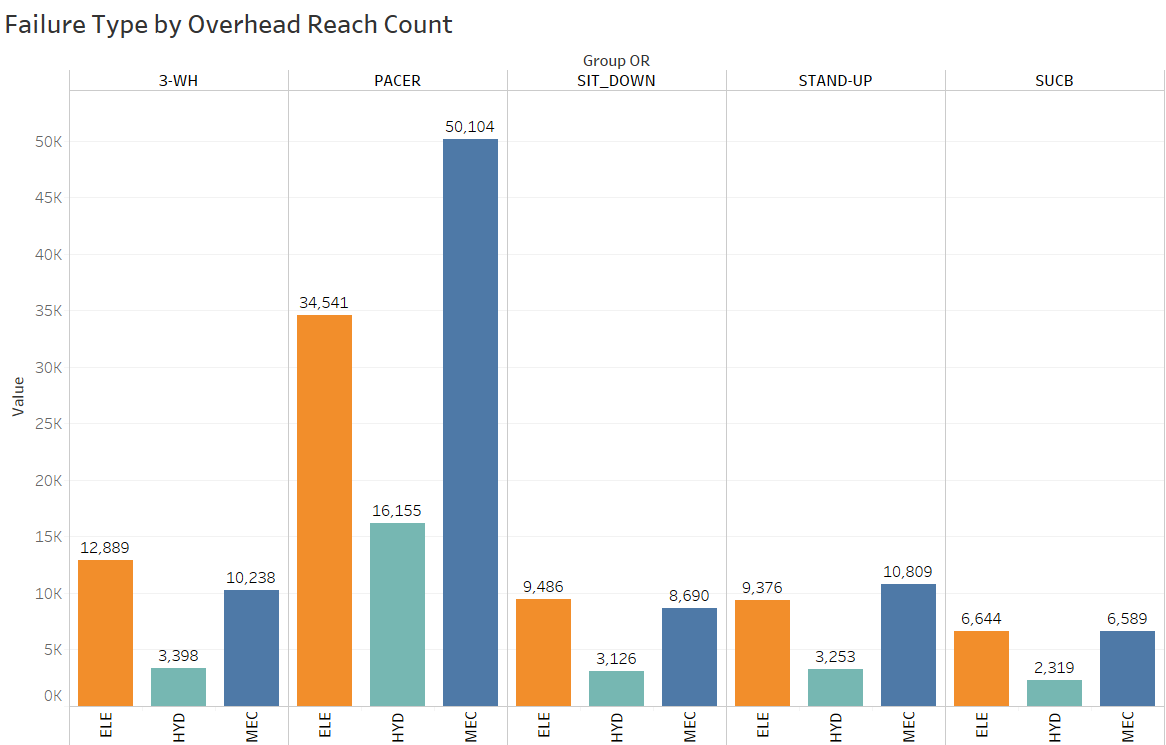
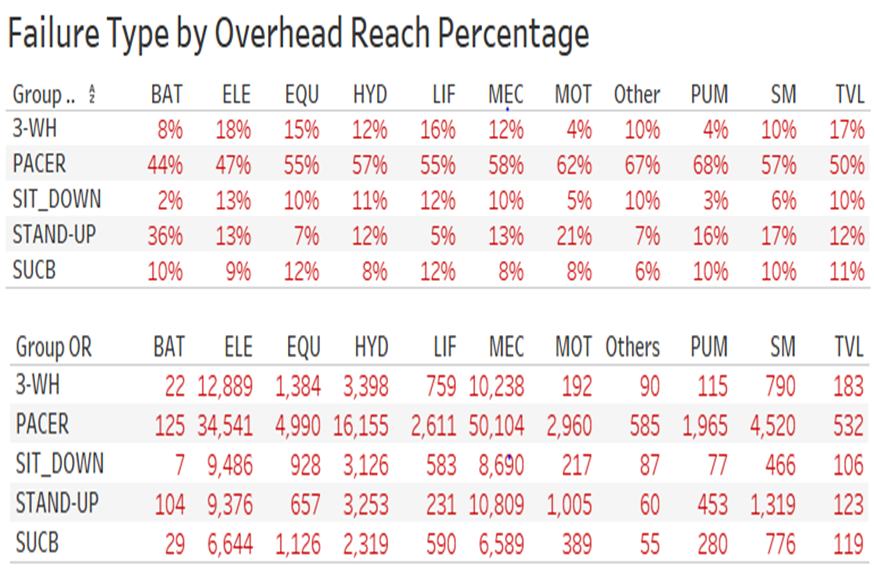
  
 Figure 24

Figure 24 provides a preliminary understanding of the distribution of the failure types and their contribution to the cost of the trucks.

It is obvious that types Mechanic (MEC), Electric (ELE), and Hydraulic (HYD) have significantly higher costs, accounting for 41% (33.91 million), 31% (25.70 million), and 17% (13.53 million) of the total cost, respectively.

After identifying the three main failure types (in terms of cost contributions), the trucks were broken down by Overhead Reach groups in Figure 25. A bar plot of the raw occurrence counts of the three main failure types within each Overhead Reach group is shown below. A table of counts and corresponding percentages is shown after.

  
 Figure 25



It could be concluded that PACER type has significantly more failure occurrences, comparing to all other Overhead Reach types. It accounts for more than half of the failure occurrences for all but two failure types. For example, according to the above table, 57% of all hydraulic failures and 58% of all mechanical failures occurred to PACER type trucks. In addition, battery and other types of failures have the least frequencies.

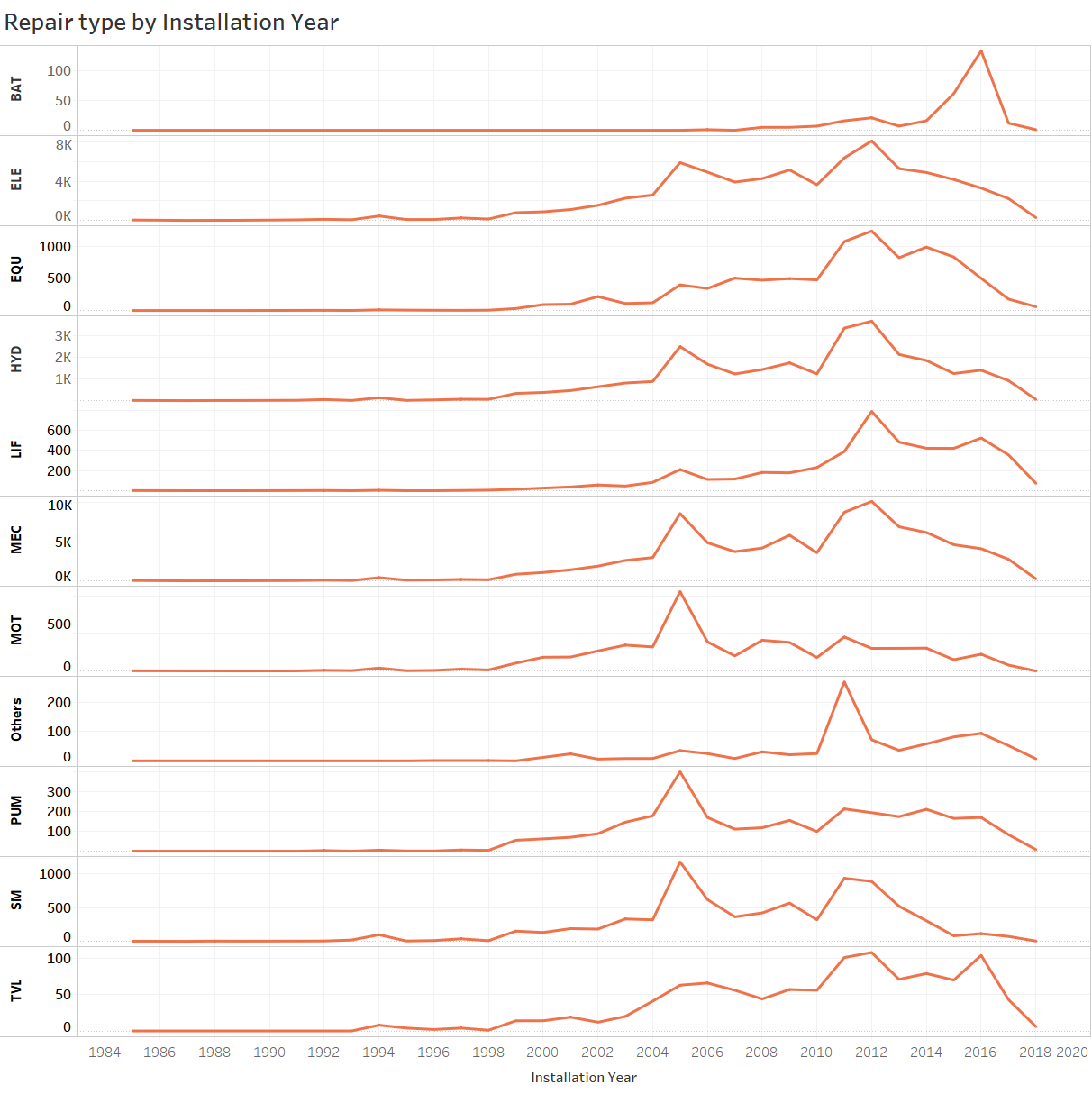
  
 Figure 26

Figure 26 shows the pattern of counts of each failure type against the installation years of the trucks. It appears that for trucks installed in years 2005 and 2011, most failure types experience significantly more occurrences. Also, The Battery (BAT) failure type was not introduced until 2016, which explains the peak in 2016 of the first line in the above plot.

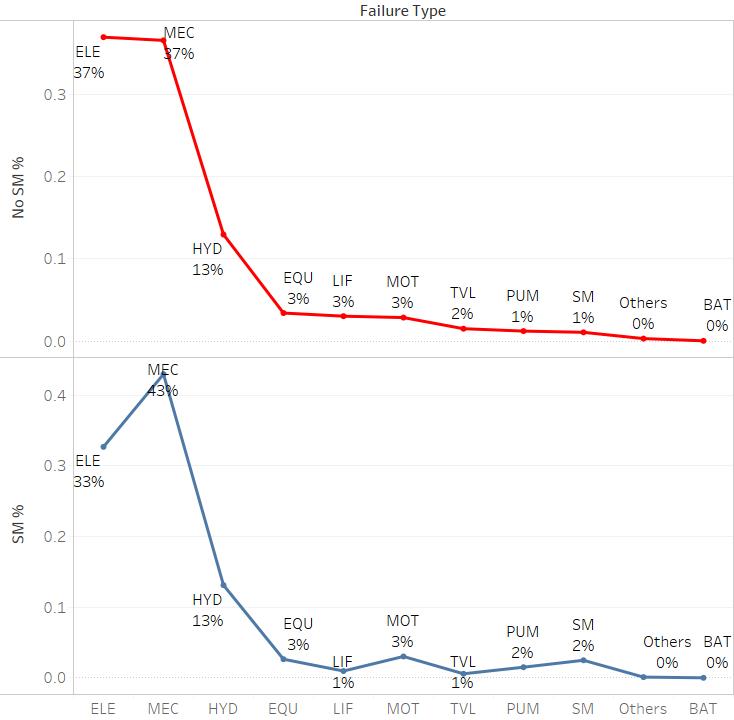
After some broad understanding of the occurrence of each failure types, some data cleaning is again required before further analysis could be conducted.

Two groups of trucks were identified based on the following criteria:

* If the last day of normal (non-SM) service happened one year or more after the last day of SM service, then the truck is assigned to group A;
* If the last day of SM service happened one year or more after the last day of normal (non-SM) service, then the truck is assigned to group B.

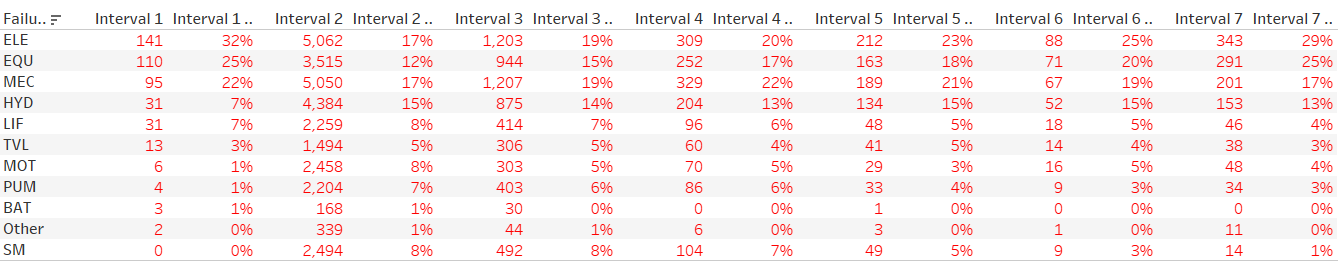
The idea is to find out the trucks which used to have an SM plan and then didn’t (group A), and the trucks that didn’t have SM plan and then did (group B).

Then, for both groups of trucks, the failures occurred for trucks not having SM plans were found out and stored in one variable, and the failures occurred for trucks having SM plan were stored in another variable. A comparison of the two variables is shown in the Figure 27.

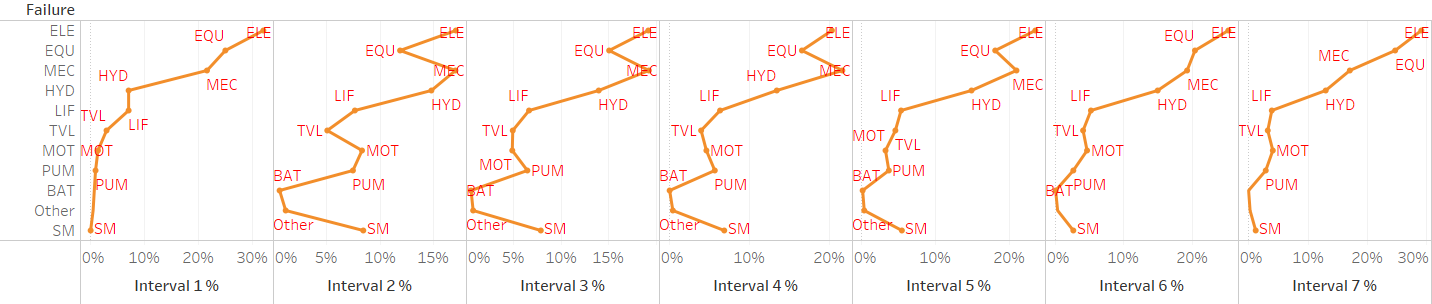
  
 Figure 27  
  
Figure 27 shows the frequencies of each repair type when a truck did not use an SM plan (Red) and when a truck used an SM plan (Blue). It could be seen that failure types ELE, LIF, and TVL occurred less frequently when SM plan was used compared to when it was not used.

A two-sample Kolmogorov-Smirnov Test (KS-Test) was conducted to see whether the distributions of the percentages of the two variables (shown in the plot) were the same, and the conclusion was positive (p-value of KS-Test was 0.9251).

The next question of interest is whether decreasing SM intervals affect the frequencies of certain failure types. Using a similar logic, failures of trucks in each SM interval group were extracted and stored in seven variables accordingly (e.g., SM\_group1\_failure, SM\_group2\_failure, etc.). Counts and corresponding percentages of the seven variables are shown in the frequency table below.



If interval group 1 is not considered (as it basically consists of outlier entries), it could be concluded that failure types ELE and EQU occur less frequently by decreasing SM time interval. Pairwise KS-tests were also conducted on the columns of percentages, and the results were all insignificant, meaning that the percentages all have the same distribution. The plot below confirms this conclusion by exhibiting similar patterns for the seven lines.

  
 Figure 28

**Optimal Service Time Interval Prediction**

**Dataset Creation**: Once the data structure was created as above, for the trucks having both: Parts & Labor Services and SM services and also trucks which have only SM services. The final business question was prediction of Optimal Service Time Interval by minimizing cost. The dataset for this would consist of demographics of the truck which are known before-hand.   
  
Based on the various demographics of the truck which are only known before-hand, a model to predict Optimal Service Time Period was built by first minimizing cost per hour for each truck.  
  
The truck features used for this model to predict Cost per hour were:

* SM Interval Group
* Installation Year
* Customer State
* Overhead Reach
* Mast Type

Note: These were the limitations faced to predict the cost per hour due to the business problem in question.

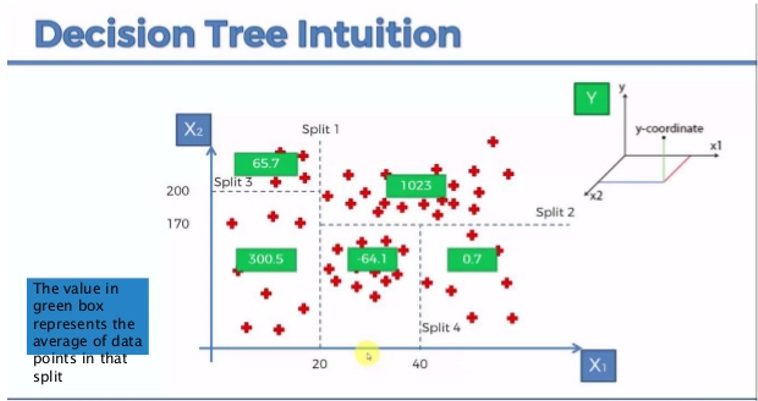
**A brief description behind the Random Forest regression algorithm:**

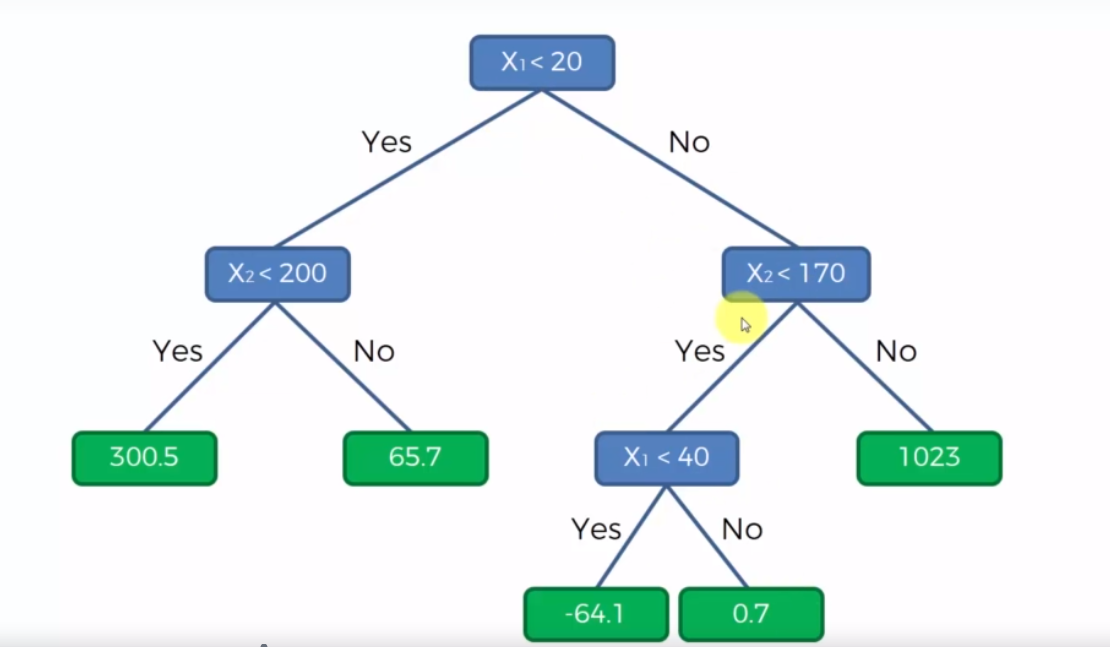
1. Pick at random k data points from the training set
2. Build the decision tree associated to these k data points
3. Choose number of trees (Ntree) wanted and repeat steps 1 and 2
4. For each now data point, make each of the Ntree predict the value of y (response variable) for the data point in question, and assign the new data point the average across all of the predicted y values.

**A brief description of how Decision trees work:** Below, are examples of two pictures how the decision tree regression algorithm makes its splits and how the child nodes for the tree are formed.

For any prediction , the algorithm takes averages of sum of squares in each terminal leaf, and assigns that value to the prediction.

The entropy measures the disorder in a dataset after a split occurs. It makes the algorithm split the datapoints by reducing the standard deviation of predictions. The standard deviation is reduced after a split occurs. The more the standard deviation decreases, the more homogenous the child nodes are after a split. Hence, the more homogenous is the data, the lower is the entropy after the split. Hence more the number of splits, more are the chances of finding parts that are more homogenous, and hence lower is the entropy. \*





\*www.superdatascience.com/pages/machine-learning

Hence, the random forest regression algorithm is better than a single decision tree as any change in the dataset does not affect all the trees much.

In other words, in building a random forest, at each split in the tree, the algorithm is not even allowed to consider a majority of the available predictors.

Suppose that there is one very strong predictor in the data set, along with a number of other moderately strong predictors.

Then in the collection of bagged trees, most or all of the trees will use this strong predictor in the top split. All of the bagged trees will look quite similar and the resulting prediction will be highly correlated. Unfortunately, averaging many highly correlated quantities does not reduce variance.

Random forests overcome this problem by forcing each split to consider only a subset of the predictors. Therefore, on average many splits will not even consider the strong predictor, and so other predictors will have more of a chance.

We can think of this process as decorrelating the trees, thereby making the average of the resulting trees less variable and hence more reliable.

**Model creation:** After the dataset was created to build the model, the basic idea used to minimize the cost per hour for a truck was scenario prediction using 7 Random Forest Regression models with 500 trees and using the “entropy” measure which reduces the standard deviation of predictions to help increase information gain. Hence, the predictor variable was cost per hour.

Each model predicts the cost per hour for all trucks falling in each of the 7 SM Interval groups respectively.

Based on clubbing the predictions of each model, the result is in the form of a 7-column matrix with cost per hour predicted for all trucks for each of the 7 SM Interval groups respectively (these are 7 different scenarios for each truck).

From this matrix, with each truck having 7 different costs per hour in each column, a minimum cost per hour was chosen along with it’s corresponding SM Interval Group (column number).

This results in a final value which tries to minimize the cost per hour for each of the truck.

An example is shown below:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cost per hour predicted from the model** | | | | | | | | |
| **If Interval 1** | **If Interval 2** | **If Interval 3** | **If Interval 4** | **If Interval 5** | **If Interval 6** | **If Interval 7** | **Minimum predicted** | **Interval Chosen** |
| 7.07 | **2.246** | 3.63 | 3.027 | 2.81 | 2.7 | 2.879 | 2.246 | 2 |
| 2.408 | 2.24 | **1.769** | 1.93 | 2.247 | 2.568 | 3.46 | 1.769 | 3 |
| 1.875 | 1.962 | 1.975 | 1.771 | 2.206 | 2.283 | **1.57** | 1.57 | 7 |
| 4.633 | 1.851 | 4.068 | 2.09 | 2.168 | 2.153 | **1.67** | 1.67 | 7 |
| 6.633 | 3.776 | 2.961 | **1.875** | 2.069 | 1.951 | 2.597 | 1.875 | 4 |
| 3.946 | **1.546** | 3.656 | 1.961 | 2.642 | 2.456 | 2.054 | 1.546 | 2 |

Note: These are not actual results, just shown for explanatory purposes.

Like the example above, the model predicts the minimum results (shown in green) and chooses the appropriate SM Interval Group for that minimum result.

This model provides a 60% reduction in cost per hour in the test set by minimizing the cost per hour and assigning the appropriate SM Interval Group to the respective trucks using this model, which meets the main business objective of minimizing the cost per hour for each truck.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1:** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **Total** | 404 | 364 | 82 | 217 | 61 | 230 | 396 |
| **Percentage** | 23% | 21% | 5% | 12% | 3% | 13% | 23% |
|  |  |  |  |  |  |  |  |
| **Table 2:** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| **Total** | 62 | 161 | 471 | 626 | 311 | 74 | 49 |
| **Percentage** | 4% | 9% | 27% | 36% | 18% | 4% | 3% |

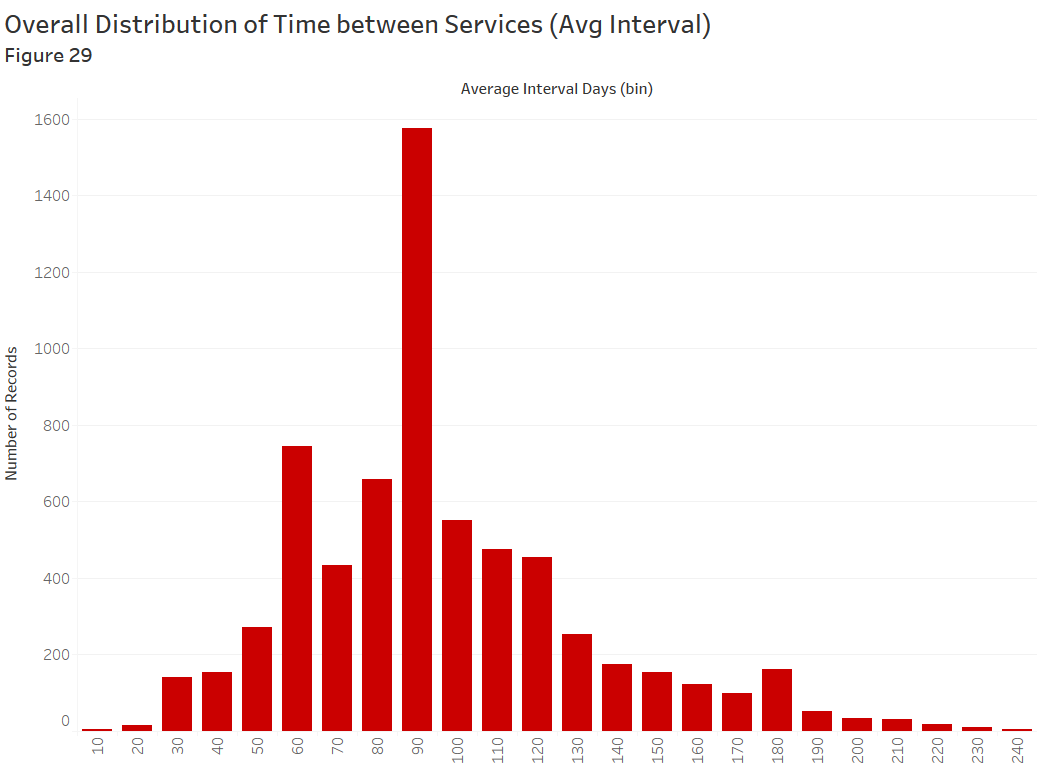
Table 1 shows count of services within each SM Interval based on the predicted model tested on the test dataset while Table 2 shows the actual count within each SM interval group based on the test dataset.

Based on table 1, which is the only possible validation measure, the model does reasonable prediction of SM Interval Groups based on the reduced costs as more than 50% of the data is assigned to Interval groups 2-6, which are the groups with most regular services.

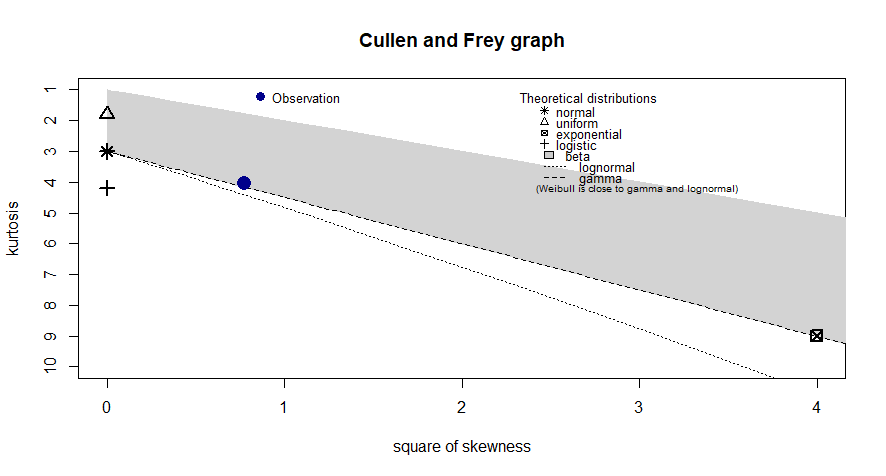
This seems realistic as Interval group 1 is for trucks with just 1 service to date, and group 7 is the group with too many days between services (contains outliers).

**Conclusions and Recommendations**

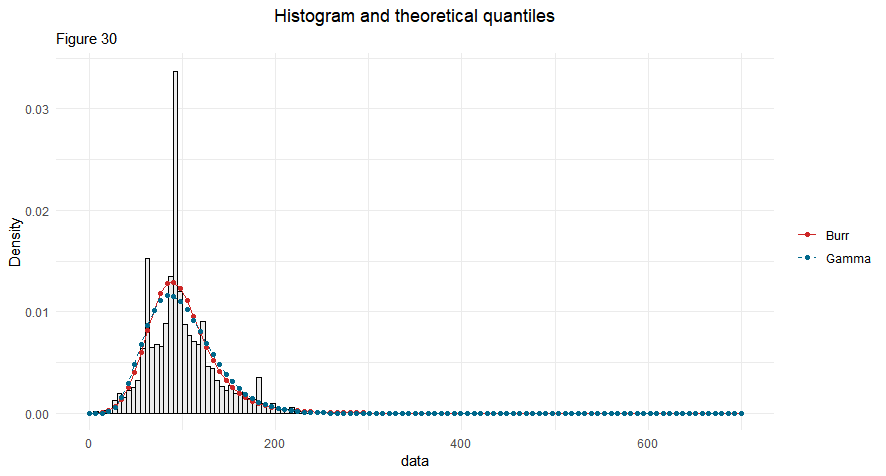
**Conclusion:** For the conclusion, some analysis was re-done leaving out Interval groups 1 and 7 to remove the effect of outliers.   
**1.** **Average SM Interval (Only SM Trucks):** This analysis was re-done without Interval groups 1 and 7. The plot for the distribution of Average Intervals for trucks is shown in Figure 29 below:

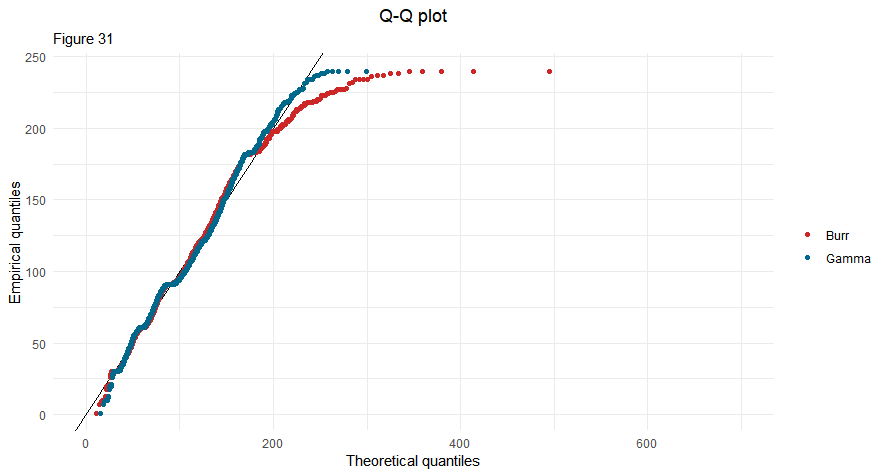


After removing all the outliers, the best fitting distribution to this data was the Gamma distribution.



The Cullen and Frey graph now shows that the Average Interval data now lies on the Gamma distribution line. Below Figures 30 and 31 compare the Burr and Gamma distributions:





Figures 30 and 31 agree to the conclusion that the Gamma distribution now fits the data almost perfectly.

**2.** **Optimal Service Time Interval Prediction (Only SM Trucks):** After re-running the model on this new dataset excluding Interval groups 1 and 7 reduces the cost by 54% on the test dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3:** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **Total** | 0 | 775 | 216 | 229 | 135 | 288 | 0 |
| **Percentage** | 0% | 47% | 13% | 14% | 8% | 18% | 0% |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 4:** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| **Total** | 0 | 161 | 471 | 626 | 311 | 74 | 0 |
| **Percentage** | 0% | 10% | 29% | 38% | 19% | 5% | 0% |

Table 3 is like table 1 which is the result from model predictions on the test dataset, while table 4 is the test dataset after excluding Interval groups 1 and 7. Based on the table 3, which is the only possible validation measure, the model does reasonable prediction of SM Interval Groups based on the reduced costs as around 47% of the data is assigned to Interval groups 2, which is the group with Interval of [0,30) days.

**Suggestions:** One suggestion to Raymond for removing the effect of some outliers in the case of services starting again after a gap of few years due to a change in client, or irregular services would be to assign a new Truck ID to those trucks when the SM services restart after a gap of too many days, as this would count that truck as a new truck and the effect of outliers would be negated.

**Acknowledgements**

Special Thanks:  
Our advisor: Dr. John Bunge  
The Raymond Corporation Team