Earthworks Excavation Project Analysis

Inayath Noon

Data Scientist 2

Assignment Overview

Tool - Python 3 on Jupyter Notebook

Libraries - numpy, pandas, matplotlib,, seaborn, scikit-learn, statsmodels

Data - 50 rows X 11 columns

- 1 Nominal Variables: Truck Name
- 2 Discrete Variables: Days (ignition), No. of Trips
- 8 Continuous Variables: Total Production, Ignition On Hrs, Utilization Hrs, Distance (Km), Speed (Km/h), Loading Time (min), Unloading Time (min), Total Emissions

Data Handling

Days of ignition above 31 for October

- Could be a technical error (might be including Nov 31st or Dec 1st)
- Field investigation on calculation required
- Whole column reduced by 3 for now

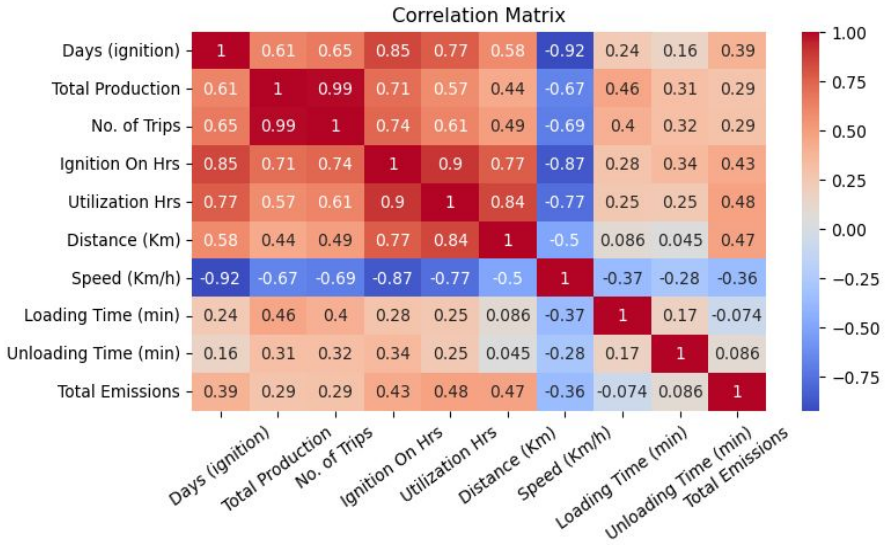
3 trucks with 0 trips

- 6% of data
- Ask onsite team the reason
- Removed from analysis for now
- Could be maintenance or breakdown, but for a whole month?

Correlation Analysis

- **Extreme multicollinearity between Total Production & No. of Trips**, we shall remove No. of Trip (if we have to build an ML model)

- **High correlation between Speed, Days (ignition) and Ignition On Hrs as well,** we shall retain them for certain analysis but remove if need be



Imputation

8 columns are completely filled, 3 require attention

1. Loading Time (min)
2. Unloading Time (min)
3. Total Emissions

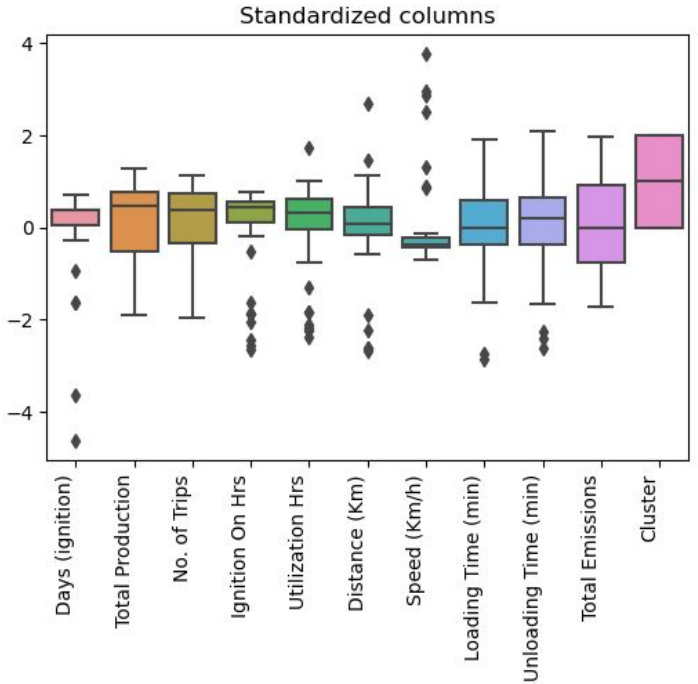
Intuitively loading and unloading could be correlated
but correlation matrix say the opposite

Loading Time (min)& Total Emissions can never be truly 0
hence we assume them random error from onsite

We shall use mean imputation. Though they are random error, conducting MICE can lead to overfitting (just 50 rows)

Outlier Detection

- Standardised numerical values and plotted as box plot to see the spread of values
- Outlier does not seem to be data errors or random error
- Very much possible the day of ignition were less for particular trucks
- Speed limit is 25 kmph, and Truck A-48 is violating it by far



Analysis - Project Progress

- Total excavation required (m3) -> 2200000
- Total excavation in October (m3) -> 874404
- Excavation completed in October (%) -> 40
- Daily excavation in October (m3) -> 28207
- No of days to complete excavation required at the october rate -> 47

Excavation is expected to completed by mid december (17th) at october rate

Analysis - Emission

Average emission from the fleet -> 8021
Lets analyse the impact of removal of 4 trucks with higher emmision (*~2x of average truck emission*) from the fleet

- Total excavation in October (m3) -> 775014
- Excavation completed in October (%) -> 35
- Daily excavation in October (m3) -> 25000
- No of days to complete excavation (at the october rate) -> 57

Even after removal of high emission trucks, we shall still be able to complete the excavation before deadline

- Emission reduced -> 62623
- Emission reduced -> 19.9 %

This decision will lead to a 20% decrease in the total emission in the next 2 months

Truck Name	Total Emissions
A-3	15392.6
A-8	15550.1
A-16	16865.4
A-23	14814.6

High Emission Trucks

Analysis - Speed Limit

Due to safety reasons trucks cannot move more than 25 km/hour speed. Since what we have is avg (not max) speed, we shall lower the threshold by 3 to be more precise

Truck Name	Distance (Km)	Speed (Km/h)	Total Production
A-20	804.96	24.630373	1207.373682
A-39	4745.59	34.654694	4376.158593
A-40	1944.76	37.693059	1226.953146
A-41	3693.43	38.546182	814.754538
A-48	2261.10	45.395546	413.233574

High Speed Trucks

- Average production of fleet = 18023
- Total Production of 5 high speed trucks = 8038
  (Half the production of an average truck)

Due to its non safe driving coupled with lower efficiency,
We have to notify this matter to the supervisor of the drivers

It could also be that these are not normal trucks but high speed, less capacity truck in safe environment with different tasks

In that case, we need to remove them from certain analysis (eg. production time) however involve them in certain analysis (eg. cost analysis)

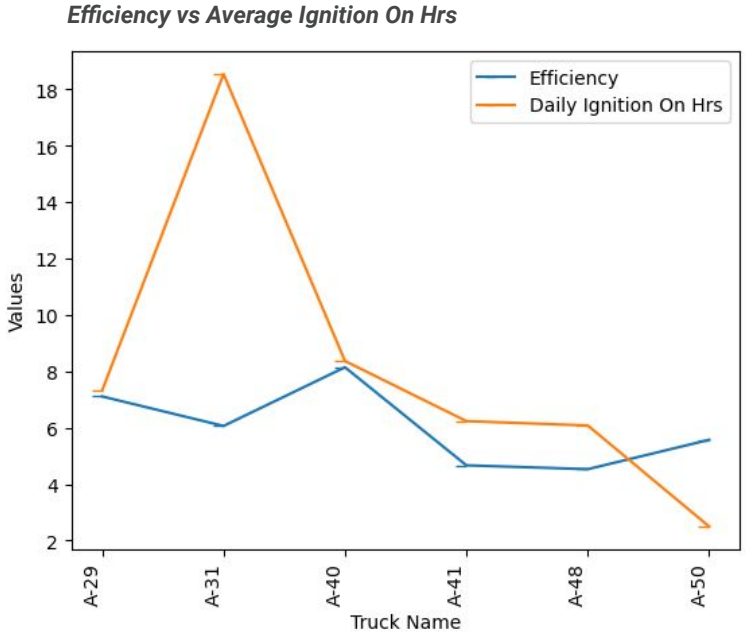
Analysis - Efficiency

We observed good correlation (0.7) between production and ignition on hrs

Let's introduce a metric

Efficiency = Total Production / Ignition On Hrs

- A20 A50 A41 A48 - maintenance report required
- The 5 trucks with the least efficiency might need maintenance
- A31 - field report required



Hypothesis Testing

We are not considering 'No. of trips' due to very high correlation with total production, to avoid multicollinearity. We shall test the significance of other variables

Null Hypothesis: There is no significant relationship between Total Production and other variables. alpha = 0.05

OLS Regression Results	coef	std err	t	P>|t|	[0.025	0.975]
const	2052.7564	2.68e+04	0.077	0.939	-5.24e+04	5.65e+04
Days (ignition)	-293.1463	698.156	-0.420	0.677	-1711.971	1125.678
Ignition On Hrs	62.7146	25.697	2.440	0.020	10.491	114.938
Utilization Hrs	8.1482	39.951	0.204	0.840	-73.041	89.338
Distance (Km)	-3.2604	3.953	-0.825	0.415	-11.295	4.774
Speed (Km/h)	36.7041	468.262	0.078	0.938	-914.918	988.327
Loading Time (min)	2245.2994	1364.227	1.646	0.109	-527.144	5017.743
Unloading Time (min)	-2927.2174	3921.090	-0.747	0.460	-1.09e+04	5041.396
Total Emissions	0.0451	0.279	0.162	0.872	-0.521	0.611

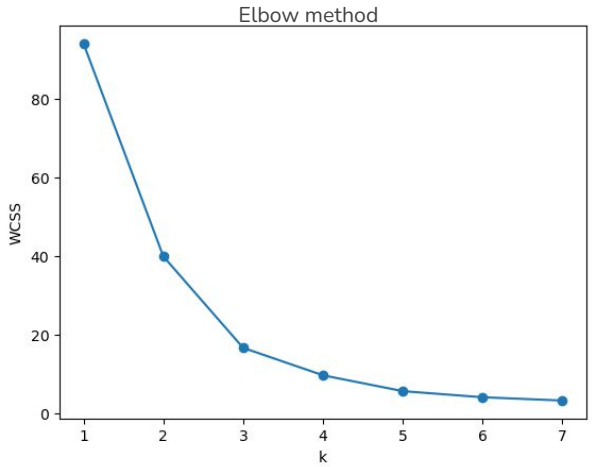
p-value of Ignition On Hrs is 0.02 (<0.05), so we reject the null hypothesis and confidently state that Total Production is directly dependant on Ignition On Hrs

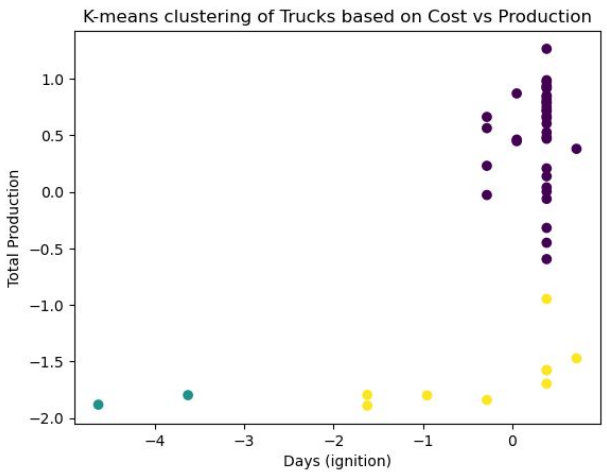
Clustering

Based on cost and production

Since cost is fixed for any truck at SAR1000/truck/day,
we can substitute days (ignition) for cost

Assumption -> per day means per Days (ignition) - note for evaluator



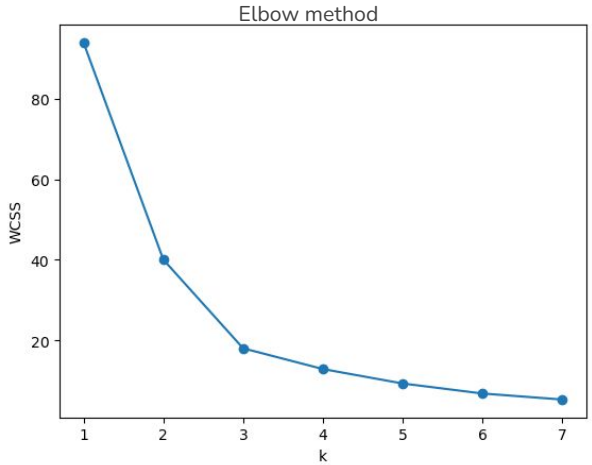


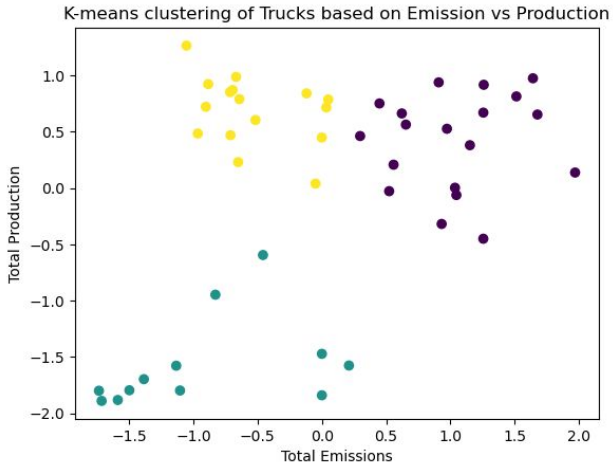
Clustering

More interesting cluster to look would be Production vs Emission

There are precisely 3 cluster ->

- LE-HP Yellow (promote),
- HE-HP Purple (discourage),
- LE-LP Green (utilise more for more data)





Cost Benefit Analysis

Increment in work due to exclusion of high emission trucks

To do the same work

Low emission truck would have to work for 38% more time than High emission trucks
Adding a fixed cost of 38% (since truck cost is same).

Other costs such as wages will also increase since more hours of work will be involved

- So basically work of 4 trucks will be completed by 5.5 trucks.
- So for 51 day 4 truck would have costed 204k SAR
- Using only Low emission truck will cost you 282k SAR (+78k)
- For 51 days 50 truck would have costed 2550k, now 2628k.

Increase to total cost of truck = 3%

Cost Benefit Analysis

Reduction in emission = (oct_em−nov_em) / oct_em

- oct_em is the total emission in october
- nov_em is forecasted reduced Emissions

nov_em = oct_em − 4*avg(removed_truck_emission) + 5.5*avg(other_truck_emission)

Percentage Reduction in Emissions = 6%

So 3% increase in total cost will result in 6% reduction in total emission

This is with the assumption that removed trucks cannot be replaced with better trucks, but the removal will reduce our cost. If it can be replaced at no extra cost, 6% emission can be reduced with no over cost.

Analysis - Utilization

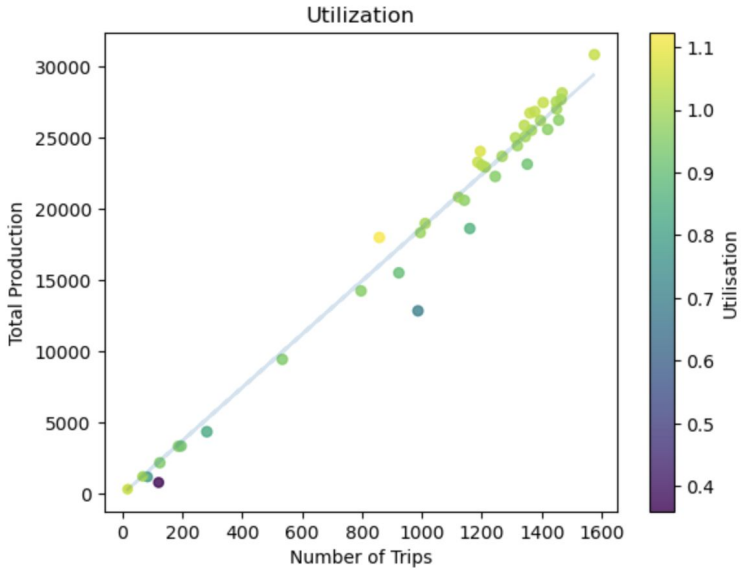
Total Production vs Utilization hr;
p value = 0.840;
so Utilization hr has no significance in Production

Lets define

Utilization = (Total Production / Total Safe Capacity)

Unsafe Trucks = Trucks with Utilization > 105%
Underutilized Trucks = Trucks with Utilization < 80%

Most truck have an optimal utilisation



Utilisation Buckets

UNSAFE TRUCKS

These 4 vehicle has come to our notice to have overloading consistently above its safe capacity

Total Safe Capacity = 22m3 × .85
= 18.7 m3

	Truck Name	Days (ignition)	Total Production	No. of Trips	Total Capacity	Utilization
5	A-6	28	24061.738060	1195	22346.5	1.076756
22	A-24	30	26742.671200	1361	25450.7	1.050764
30	A-32	30	18011.559560	858	16044.6	1.122593
44	A-48	15	413.233574	20	374.0	1.104903

LOW UTILIZATION TRUCKS

These 3 truck have untapped potential, filling less than 80% of the safe capacity overall (purple in earlier below) and they collectively do the work of ~2 trucks.

We should be able to save 1/50 (i.e. 2%) of the total cost if their potentials are tapped; also save Loading and Unloading time

	Truck Name	Days (ignition)	Total Production	No. of Trips	Total Capacity	Utilization
19	A-20	26	1207.373682	84	1570.8	0.768636
20	A-22	30	12861.948650	987	18456.9	0.696864
38	A-41	28	814.754538	121	2262.7	0.360081

Analysis - Cost Efficiency

Since cost is linked to the time a truck is running 'Days (ignition)'

- It's essential to check if each truck is cost-effective
- We're using Pareto Analysis (to find out which trucks make up the top 80% of production)

Surprisingly, our data doesn't match the typical Pareto pattern.

- Out of all 43 trucks, 25 (58%) are responsible for the top 80% of production

Cost analysis is limited with this data

- Cost is fixed for all trucks
- Variable costs (or any other costs) are not available

Step-by-Step Action Plan:

Immediate Actions (week 1):

- Address overloaded and underutilized trucks.
- Conduct maintenance for trucks A20, A50, A41, A48, and A31.
- Notify supervisors about high-speed trucks violating safety limits.

Planned Actions (week 2-3):

- Enhance the efficiency of low-utilization trucks.
- Removal or replacement of 4 high emission trucks.

Other Actions:

- Explore the possibility capturing data at day level.
- Including other feature suchs as fuel consumption, downtime, route.
- Collect data to categorise truck based on their different tasks.
- Utilise the Unloading & Loading time data that os captured.

Communication:

- Report the impact of the changes on production, cost, emissions.
- Provide regular updates to stakeholders on the progress.
- Implement additional measures based on continuous feedback.

Conclusion

This analysis covers various aspects like outlier detection, correlation analysis, imputation, and hypothesis testing. Key findings include the need for further investigation into trucks violating speed limits, potential maintenance requirements for low-efficiency trucks, and a cost-benefit analysis highlighting a potential 6% reduction in emissions with a 3% increase in total costs. The project's multifaceted approach provides actionable recommendations for enhancing efficiency, safety, and timely project completion.