**Executive Summary**

The "Rolling Stock Fleet" project aimed to analyze and optimize a vehicle fleet dataset comprising 1185 rows and 12 columns. The dataset recorded vital operational data, including average mileage, downtime, and labour hours. Significant variability and data integrity issues were addressed through extensive cleaning and transformation, ensuring reliable analysis. The project employed the Random Forest Regressor to model non-linear relationships within the dataset, achieving high predictive accuracy. Visualizations highlighted fleet composition, usage, and maintenance needs, facilitating informed decision-making. Recommendations focused on reducing fleet size, improving sustainability, and enhancing operational efficiency.

**Problem Statement**

The Rolling Stock Fleet dataset, capturing key operational data, faced significant challenges due to missing and inconsistent data, particularly in "HourMeter2020(hours)" and "Mileage2020 (km)." These issues compromised the dataset's reliability, necessitating comprehensive data cleaning and transformation. The objective was to ensure data integrity for accurate analysis, guiding fleet optimization and reducing operational costs and environmental impact.

**Data Cleaning**

Ensuring data integrity is crucial, especially for the "Rolling Stock Fleet " dataset, capturing vital operational data of a vehicle fleet. Despite its comprehensive detail, with 1,185 rows and 12 columns charting fleet usage and operational status, an average mileage of about 7,062 km, with downtime and labour hours averaging 197 and 55 hours respectively was observed. The data highlights significant variability in vehicle usage and maintenance needs, with some extreme values pointing to high mileage outliers and instances of extensive downtime or labour hours, it encountered issues like missing data and inconsistencies. An intensive cleaning and transformation process was thus imperative to prepare it for in-depth analysis. The cleaning began with identifying errors and missing data across essential columns, including "HourMeter2020(hours)," "Mileage2020 (km)," and more. These gaps threatened the analysis's reliability, demanding strategic interventions.

To address these challenges, the dataset underwent several corrective steps. The "HourMeter2020(hours)" column, with over 69% missing data, was removed to maintain the dataset's integrity. The next focus was imputing missing values in other columns, adopting a blend of logical assumptions and advanced statistical methods. For example, zeros replaced missing values in "DowntimeHours2020(hours)" and "Mileage2020 (km)" for certain conditions, while the K-Nearest Neighbors (KNN) method refined the imputation for other missing mileage data. These steps ensured the dataset's completeness and reliability, paving the way for accurate and insightful analysis.

**Data Transformation**

In transforming the data for analysis, the skewed distribution of "DowntimeHours2020(hours)" and "LaborHours2020(hours)" was addressed through a logarithmic transformation, aiming for normalization to enhance visualization readiness. These logarithmic features also indicate a strongly positive correlation of 83%, highlighting an important relationship between both features (see Appendix G & H). Also, one-hot encoding was applied to categorical variables like equipment category and service group, which is essential for modelling the fleet dataset. Furthermore, the consolidation of secondary CO2 emissions data from 1995 to 2019 (Government of Canada, nd) involved meticulous loading, cleansing, and vertical concatenation, culminating in a dataset rich in historical CO2 emissions insights. A left-join merge strategy, focusing on 'Make' and 'Year', ensured comprehensive integration without compromising data integrity, facilitating a broader analytical scope.

Outlier analysis revealed variances in Downtime Hours, Labor Hours, and Mileage, attributed to the fleet's operational diversity and the unique demands of specific vehicles. These outliers, indicative of genuine operational nuances or specialized vehicle usage, were preserved to maintain the dataset's reflection of the full operational spectrum. This decision underpins the dataset's comprehensive portrayal of operational realities. Following rigorous cleaning and imputation, the dataset emerged streamlined, with 1,185 rows and 11 columns primed for in-depth, accurate analysis, marking a preparation phase that advocates the commitment to data integrity and analytical reliability.

**Assumptions**

The missing values of the DowntimeHours2020 and LaborHours2020 were assumed to be zero due to no downtime or labour hours, also, an assumption was made that the heavy equipment does not accumulate mileage due to no presence of mileage for the heavy equipment. Missing mileage with zero downtime and labour hours replaced to zero on the assumption that it was not used that year. Certain assumptions were also made in creating the CO2 emissions column, the average emissions for the vehicles were created using the aggregate of the ‘year’ and ‘make’ from the secondary dataset.

**Model Selection**

The dataset's non-linear traits led to using the Random Forest Regressor, a model capable of navigating complex, non-linear relationships between variables and targets. This selection diverges markedly from linear approaches, such as linear regression, which assume linear interactions and thus fall short of capturing the dataset's subtleties. Despite neural networks' ability to model non-linear dynamics, their need for extensive preprocessing and tuning and a tendency towards opacity and overfitting rendered Random Forest a preferable choice. Impressive metrics underscored its efficacy: a Mean Squared Error of 14.5, an R² Score of 0.98, and a Mean Absolute Percentage Error of 0.44, attesting to its precise fit and high predictive accuracy.

Hyperparameter tuning via GridSearchCV refined the model's performance, identifying an optimal configuration of {'max\_depth': 10, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 50} that significantly improved model effectiveness. When evaluated on the test dataset, this tuned model achieved notable results, with an MSE of 17.29408753637539, an R² Score of 0.9876795684507761, and a MAPE of 0.47%. These outcomes demonstrate the model's exceptional accuracy and highlight the successful hyperparameter strategy, particularly the balanced approach to model complexity and performance optimization.

**Visualization**

**Fleet Composition by Service Group (Appendix A):** This graph clearly shows how the fleet is distributed across different service groups, highlighting engineering service as the area with the most resources and human resources service as underserved areas.

**Fleet Composition by Department within Engineering Services (Appendix B):** By breaking down the fleet within the Engineering Services department, this graph shows where resources are focused within a specific segment of the operation. It is helpful for department-level planning and management, helping to identify if certain areas might be over or under-resourced relative to their operational demands.

**Fleet Composition by Equipment Category (Appendix C):** This visualization reveals the fleet's diversity and types of equipment, with light duty being the highest. This is crucial for maintenance planning, procurement, and operational strategy.

**Aggregate Total Mileage by Service Group (Appendix D):** Total mileage indicates the activity level and potential wear on vehicles within each service group. Engineering Service has the highest, followed by Parks and Recreation. This signals a need for closer maintenance monitoring or fleet renewal efforts to avoid increased downtime or higher repair costs.

**Aggregate Total Labor Hours by Service Group (Appendix E):** High labour hours point to areas with intense maintenance demands or operational inefficiencies. The graph shows that Engineering Services is the highest.

**Aggregate total downtime hours by service group (Appendix F):** Total downtime hours are a critical indicator of operational disruptions. This graph shows that Engineering Services is the highest.

**Recommendations**

Understanding fleet usage and identifying opportunities to minimize fleet numbers necessitates a detailed dataset on operations, environmental sustainability, and operational costs. A dataset focused on model-specific fuel consumption and estimated CO2 emissions for new light-duty vehicles in Canada would significantly augment our analysis, providing insights into vehicle efficiency and environmental impact. This information enables strategic decisions aimed at optimizing the fleet by taking into account each vehicle's efficiency and environmental imprint, so contributing to the overall goal of fleet optimization.

The necessary data for this analysis can be sourced from the Government of Canada's open data portal, offering extensive details on fuel consumption and emissions across different vehicle models over time. Access to this repository of information ensures that the analysis is bolstered by up-to-date and accurate data on vehicle performance, enhancing the understanding of their environmental impact. Furthermore, downsizing the fleet promises considerable cost savings, reduces carbon emissions, boosts fleet utilization, and improves operational efficiency, aligning with environmental conservation efforts and preparing the fleet for future operational challenges and regulatory changes.

To effectively reduce the fleet size, a proposition of starting with an analysis to identify underutilized vehicles with high downtime and labour hours relative to mileage indicates inefficiency and elevated maintenance costs. Targeting these vehicles for removal and an age analysis to highlight older, more polluting vehicles lays the groundwork for a phased replacement strategy. Newer, more efficient models can lower emissions and improve operational efficiency as shown in the correlation matrix, the negative correlation between the year and vehicle emission (See Appendix G). Emphasizing fuel efficiency, vehicle maintenance, and operational practices will foster a more sustainable, cost-efficient fleet.

While fleet reduction harbours numerous advantages, it also presents potential challenges, including operational difficulties during peak demand, accelerated vehicle wear, and impacts on employee morale. To mitigate these risks, enhancing flexibility through partnerships with rental services for peak times, continuous monitoring to adjust strategies, and developing contingency plans are crucial. These strategies ensure the fleet remains adaptable, maintaining service levels despite fleet size adjustments and unforeseen demands.

To achieve sustainability and environmental goals, a targeted reduction in CO2 emissions across the 'light duty' fleet is strongly recommended. Key recommendations include Engineering Sanitation Operations to aim for a 25.3 million gram reduction, Engineering Streets and Sewers Operations to target reductions of 23.7 million grams each, Park Board Planning and Operations to work towards a 23.6 million gram cut, and Real Estate & Facilities Management to strive for a 20.2 million gram decrease. By following these recommendations, operations can be steered toward greater sustainability and efficiency, aligning with our environmental goals.

**References**

Government of Canada. (n.d). *Fuel consumption ratings.* OpenCanada. <https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64/resource/2309538b-53d1-4635-a88e-e237bfcef7a2>

**Appendix**

**Appendix A**

A graph with numbers and a number of pieces of equipment

Description automatically generated

**Appendix B**

A graph with different colored bars

Description automatically generated

**Appendix C**

A graph of different colored squares

Description automatically generated

**Appendix D**

A graph with a number of miles

Description automatically generated with medium confidence

**Appendix E**

A graph showing a number of workers

Description automatically generated with medium confidence

**Appendix F**

A graph with numbers and a bar

Description automatically generated with medium confidence

**Appendix G**

A screenshot of a computer screen

Description automatically generated

**Appendix H**

A group of blue and white graphs

Description automatically generated