Exploring CO_2 Emissions from Automobiles

Andrew Mashhadi | STAT411

Introduction

Carbon dioxide (CO_2) emissions is often recognized as one of the main drivers in global climate change.

Countries are urgently trying to reduce their annual emissions to prevent any further impacts of climate change.

Automobiles significantly contribute to the total annual CO_2 emissions each year.

Introduction

In this project:

- I used PCA and Factor Analysis with multiple linear regression to explore emissions data to bring out strong patterns between various vehicle attributes and CO_2 emissions.
- I attempted to characterize the key factors affecting emission levels.
- Examined the overall fit of each model, assessed the corresponding predictive performance, and investigated the relationship between the main factors and the expected CO_2 emissions.

Data

- Canadian Government's official open-access website. [2]
- Collected over 7 years.
- Over 7000 observations and 10 original variables.
- Each observation is associated with an independent automobile and includes a variety of vehicle attributes with an associated CO_2 emission (measured in grams per kilometer).
- Original vehicle attributes consisted of: *Make, Model, Vehicle Class, Engine Size, Cylinders, Transmission, Fuel Type, City Fuel Consumption, Highway Fuel Consumption*.

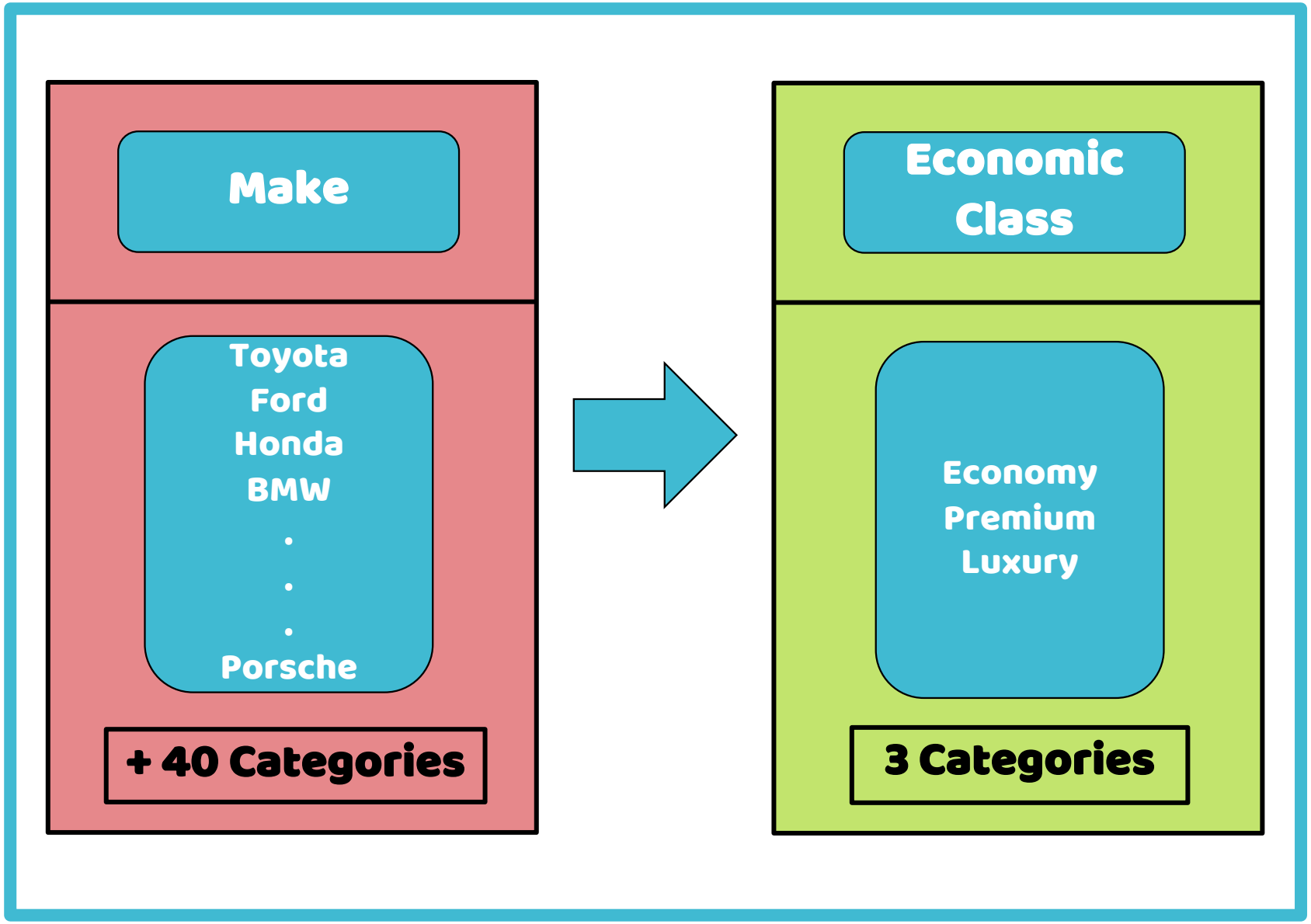
Data Cleaning and Feature Engineering

- Although we have 9 explanatory variables and 1 response variable (CO_2 Emissions), many of the categorical features contain a plethora of categories.
- This vastly increases the dimension of the feature space when dummy coding is applied to the variables.
- Most of these original categorical variables were split up, or recategorized, in an attempt limit the number of categories from a single feature, and to combine any categories based on general commonalities.

Data Cleaning and Feature Engineering

The original variable "Make" was replaced with "Economic Class" since it had too many categories.

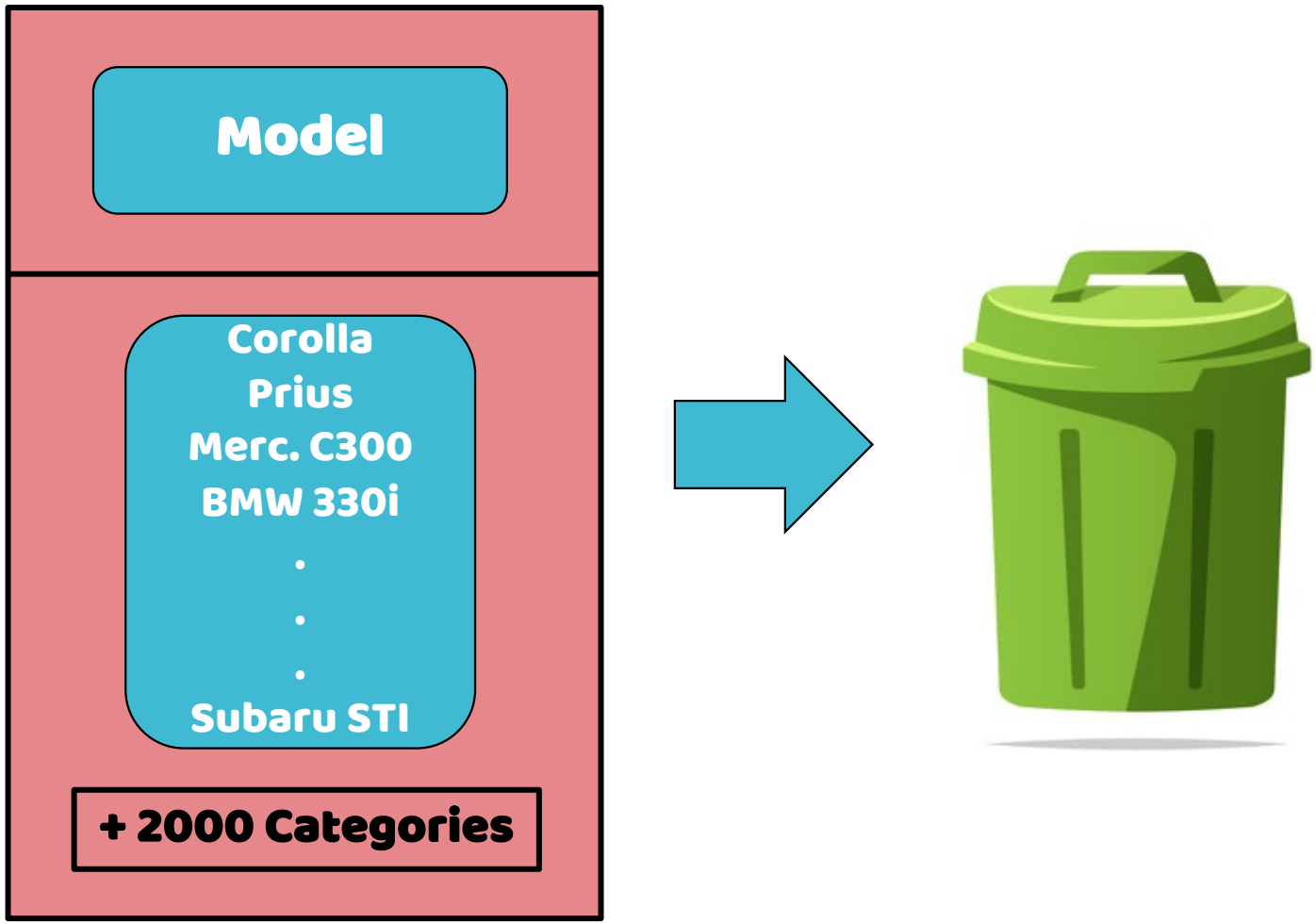
Note that the new economic class variable also has a natural order to it.



Data Cleaning and Feature Engineering

The original variable "Model" had over 2000 different categories.

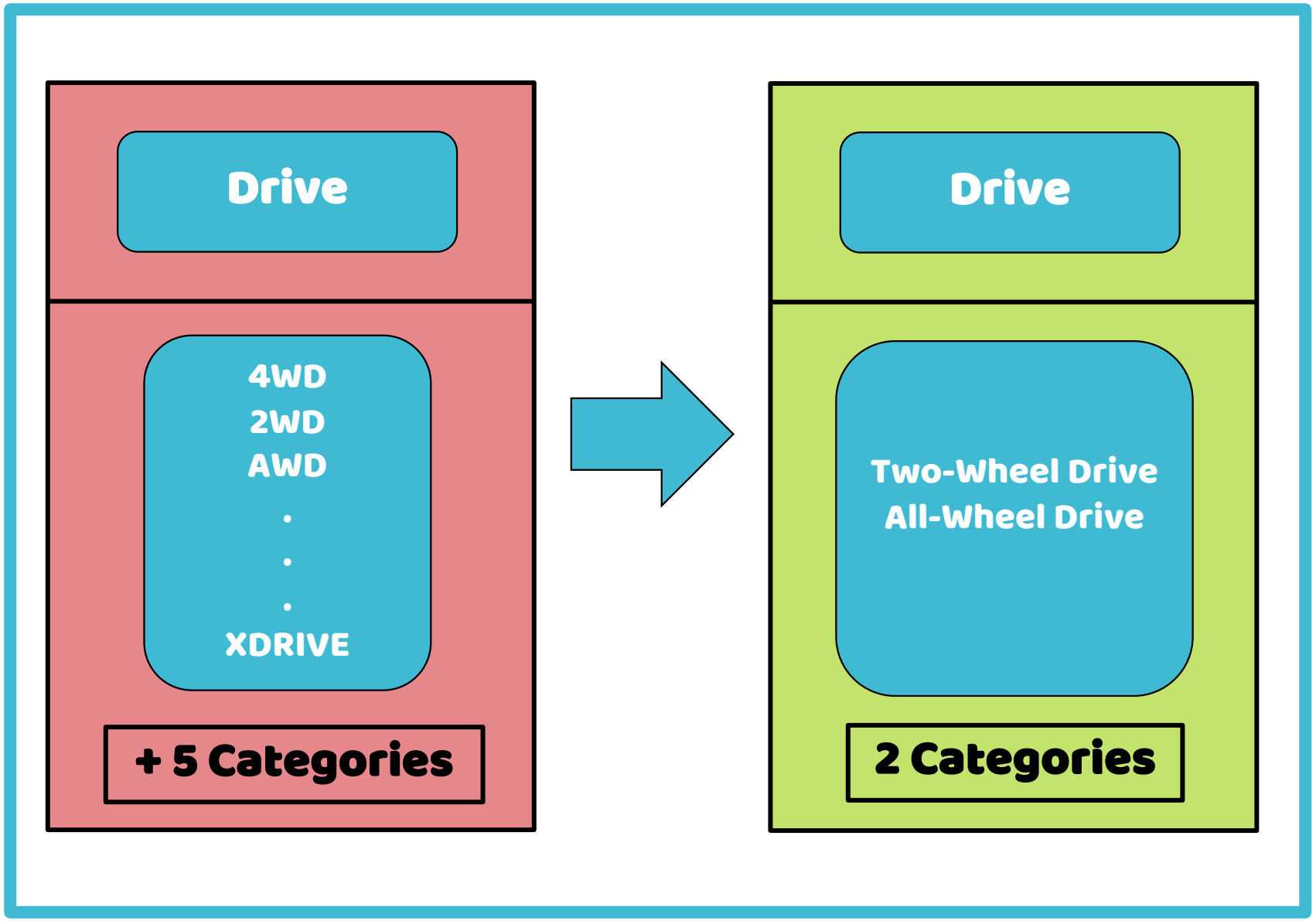
I removed this feature from consideration since there would simply be too many categories compared to the number of observations.



Data Cleaning and Feature Engineering

The original variable "Drive" had over 5 different categories, and many of which meant the same thing.

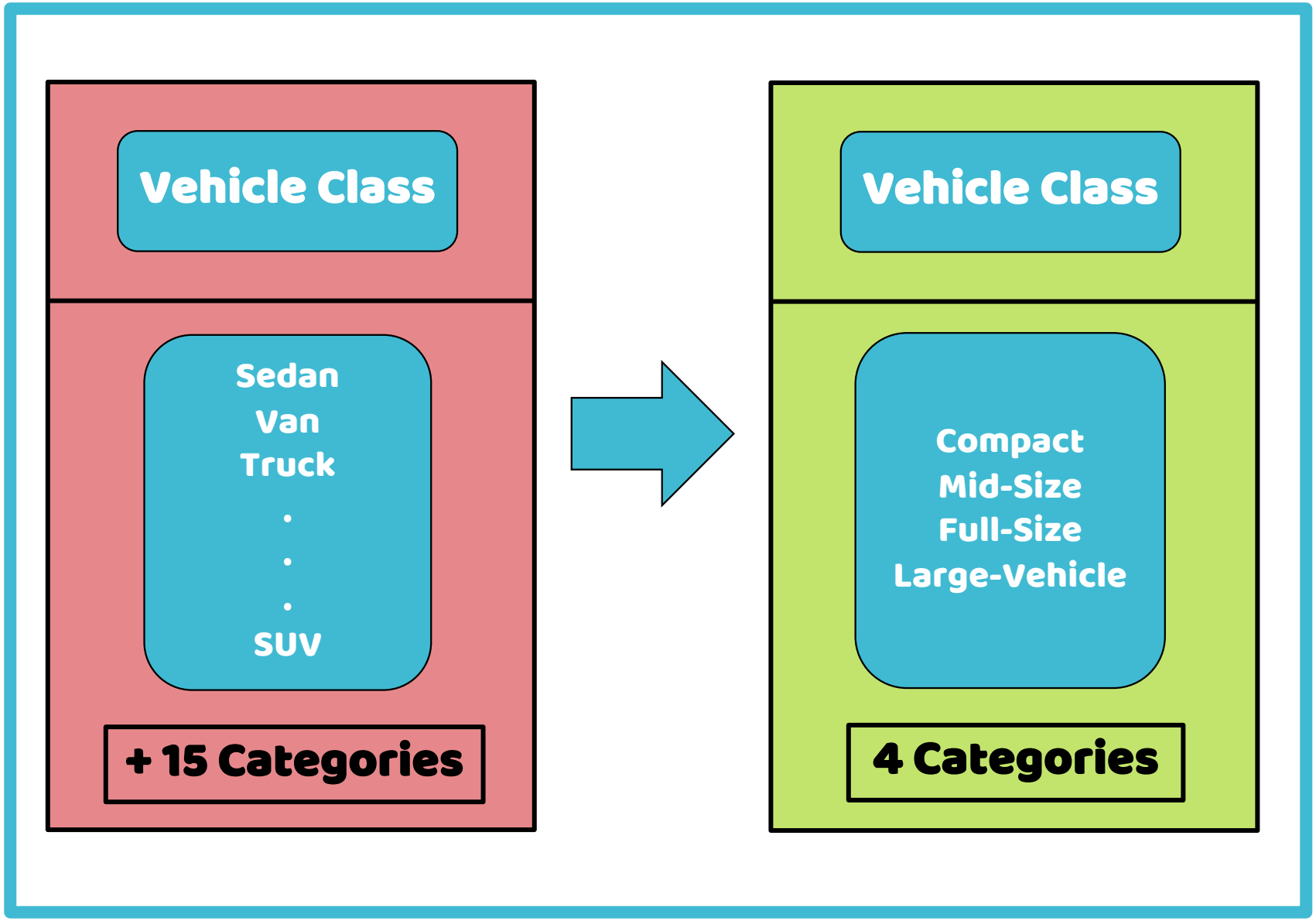
I re-categorized this variable to have 2 categories.



Data Cleaning and Feature Engineering

The original variable "Vehicle Class" had over 15 different categories, and many of them were closely related.

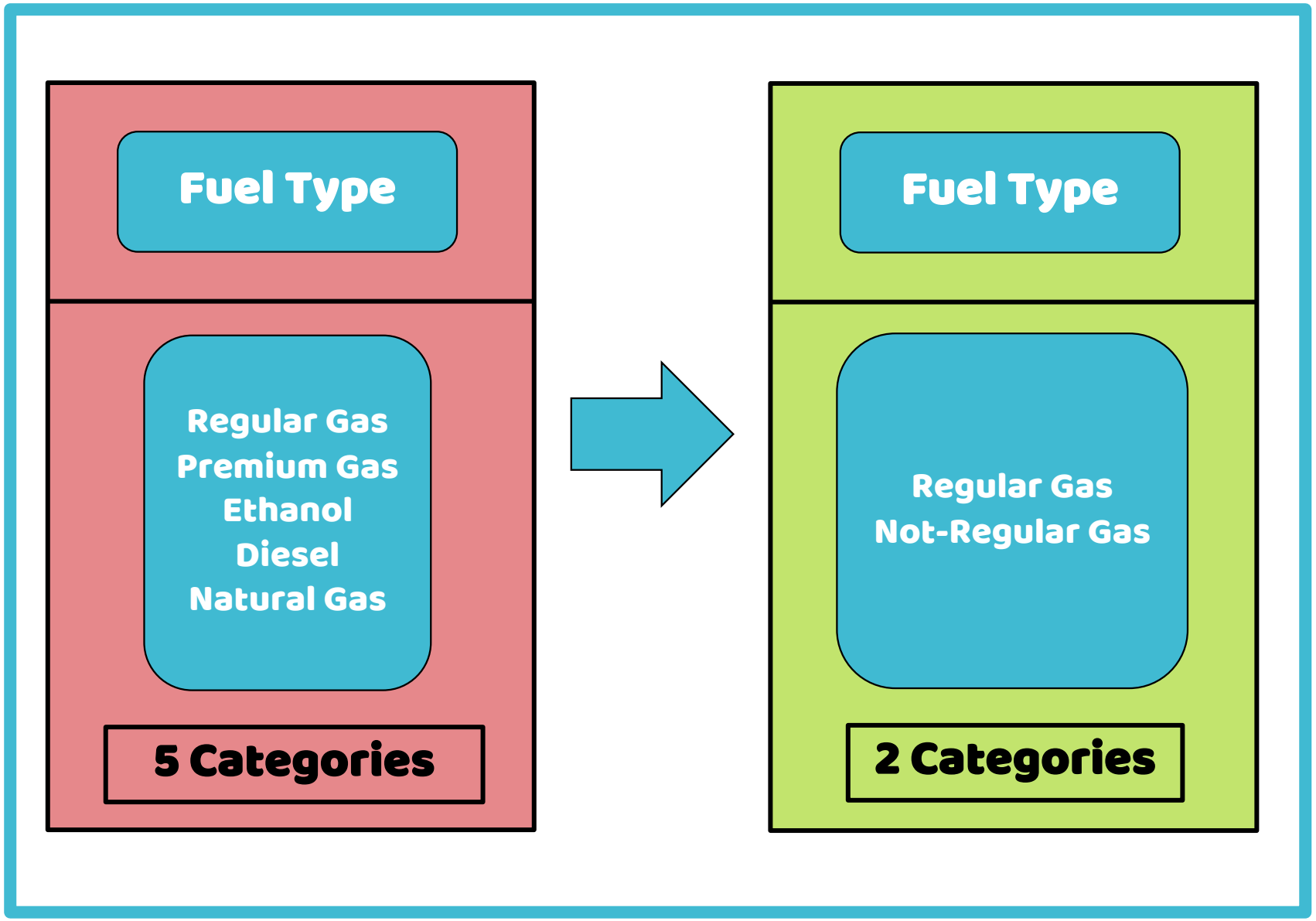
I re-categorized this variable to have 4 categories.



Data Cleaning and Feature Engineering

For simplicity, I re-categorized the "Fuel Type" variable as well.

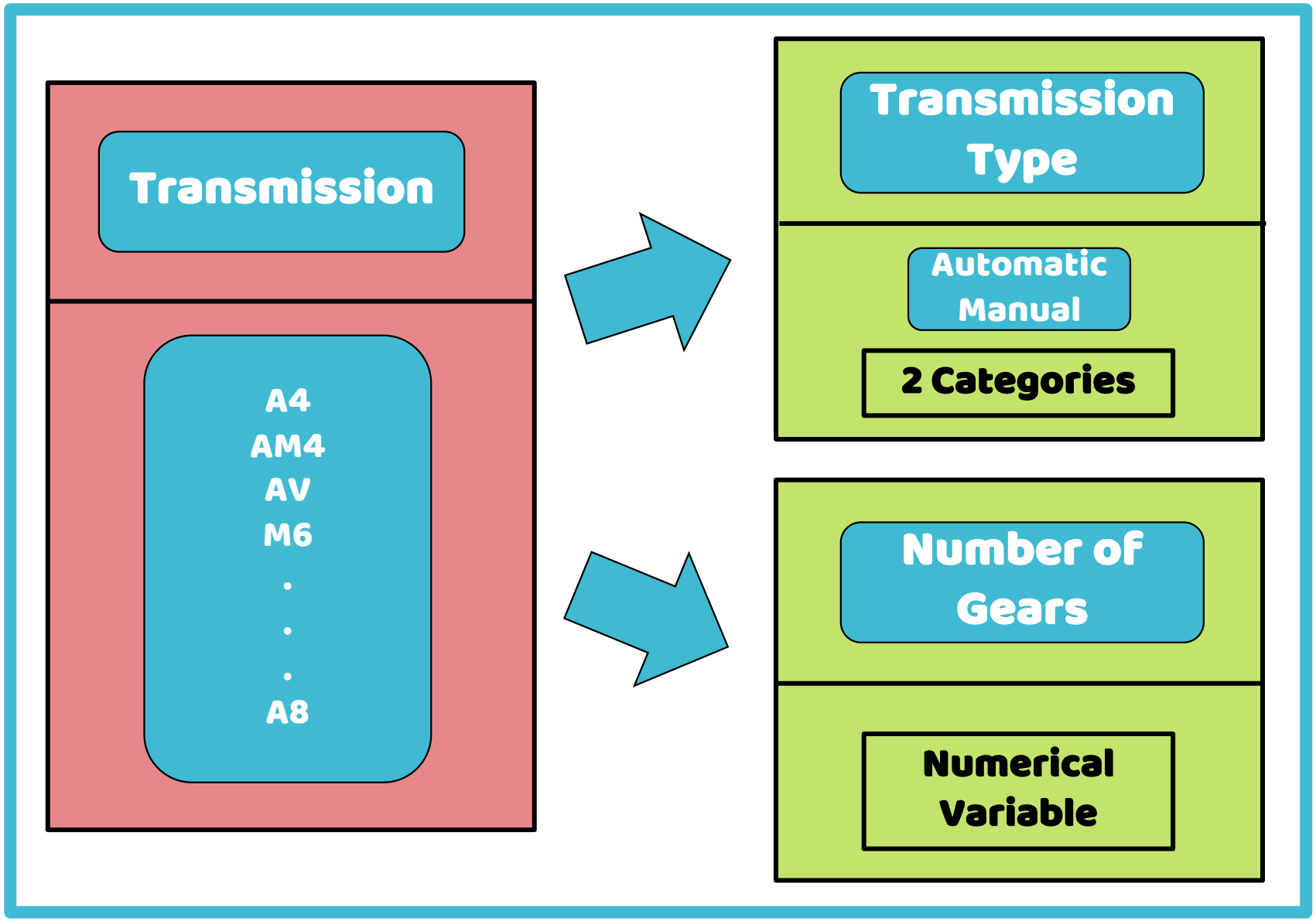
Each observation was grouped within a "Regular Gasoline" category or a "Not-Regular Gasoline" category.



Data Cleaning and Feature Engineering

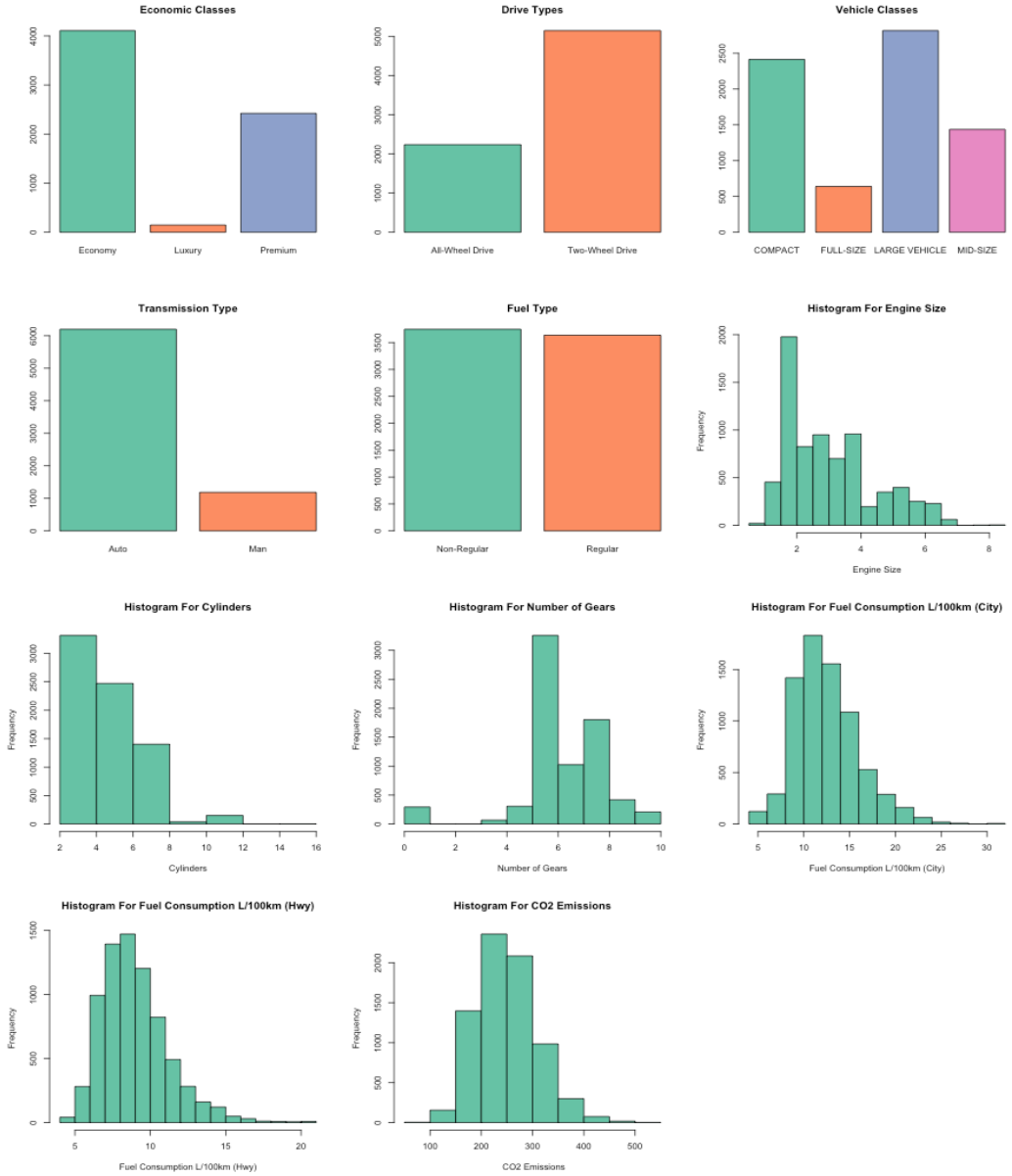
For "transmission", each observation was given a specific acronym that effectively described the type of transmission and its number of gears.

Therefore, I split this variable into two new variables: "Transmission Type" and "Number of Gears"



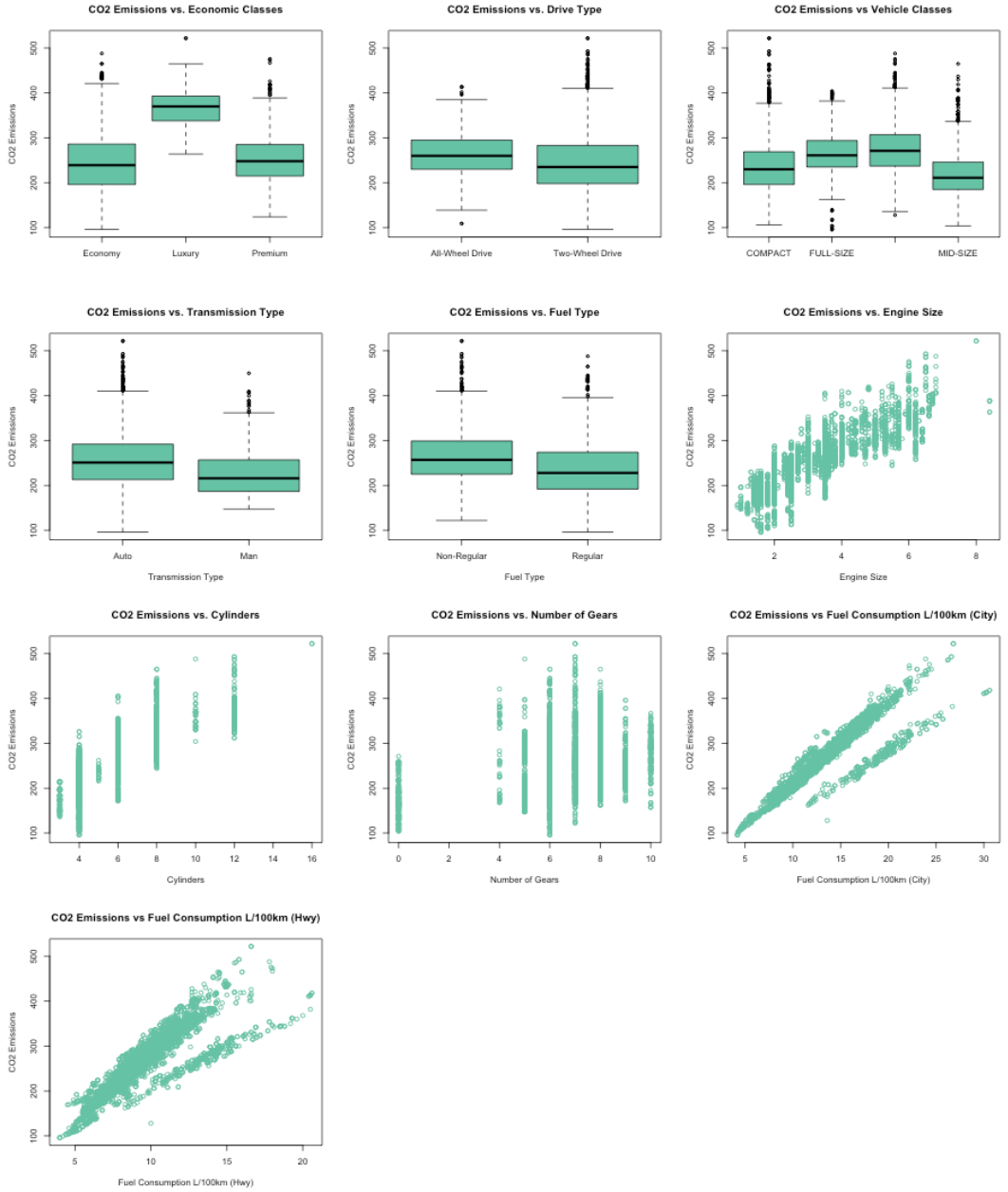
Exploratory Data Analysis

- I explored the one-way distributions for the numerical variables and the one-way frequency tables for the categorical variables.
- Notice that all the continuous variables except for Engine Size and Cylinders have approximately normal distributions, while the variables Engine Size and Cylinders seem to have right (positive) skewed distributions.
- All the cleaned categorical variables demonstrate large sample sizes (n > 100) in each category.
- CO_2 Emissions clearly demonstrated normality with an estimated mean of about 250 g/km and an estimated standard deviation of about 59 g/km.



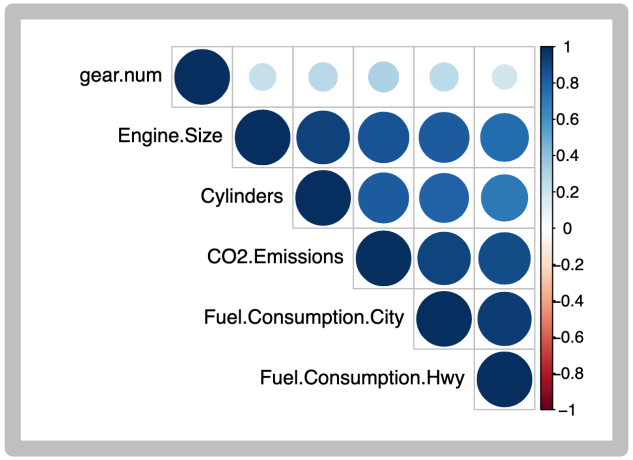
Exploratory Data Analysis

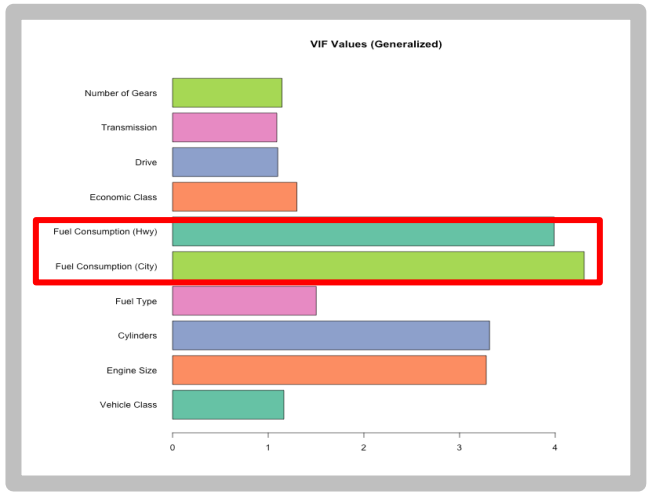
- I explored the two-way relationships between each vehicle feature and the CO_2 Emissions
- Box-plots were used for the categorical variables to illustrate potential differences in the distribution of CO_2 emissions when conditioned on each category
- Scatter-plots of the CO_2 emissions against the explanatory variable were used for the numerical features.
- All categorical variables indicated significantly different means (with similar distributions) for each category of the associated vehicle feature.
- All numerical features except for "Number of Gears" demonstrated a clear, positive, linear relationship with the CO_2 emissions.



Exploratory Data Analysis

- I investigated the relationships between the cleaned variables in the dataset.
- All VIF values are approximately less than or equal to 5, so I had no reason to believe any negative effects from multicollinearity would impact the analyses or models.
- The correlation plot indicated that most of the numeric variables have a relatively large correlation with CO_2 Emissions and with each other.
- PCA and Factor Analysis should help alleviate some of the negative effects that may arise from collinearity when applied before modeling.





Methodology

- Constructed a training and testing set with an 80/20 split of the original data.
- Since many of the variables used are not in the same units, I scaled both sets using the estimated means and standard deviations from **only** the training set.
- Applied PCA to the automobile attributes in the training data to reduce the dimensions of the data while minimizing information loss.
  - Investigated the loadings of the first few components.
  - Based on the location of the "elbow" in the corresponding scree plot and the cumulative variance explained by the principal components, I choose an optimal number of components.
  - Linearly transformed the training data into principal components that were used to fit a multiple linear regression model.
  - Linear model assumptions were checked, residual diagnostics were examined, and the significance of each coefficient was tested.

Methodology

- I also applied factor analysis to the training data, using the *Maximum Likelihood* method:
  - A *varimax* rotation was used to help "spread out" the squares of the loadings on each factor in hope of finding groups of large and negligible coefficients in any column of the rotated loadings.
  - The loadings of the first few factors were investigated and interpreted in order to characterize the key factors found among the original automobile features.
  - The location of the "elbow" in the corresponding scree plot and the cumulative variance explained by the factors was used to choose an optimal number of factors.
  - With the loadings from the first optimal number of factors, I then generated factor scores using the weighted-least-squares method shown in class:

$$\hat{\mathbf{f}}_j = (\hat{\mathbf{L}}_\mathbf{z}' \hat{\mathbf{\Psi}}_z^{-1} \hat{\mathbf{L}}_\mathbf{z})^{-1} \hat{\mathbf{L}}_\mathbf{z}' \hat{\mathbf{\Psi}}_z^{-1} \mathbf{z}_j \quad \text{for } j = 1, 2, \ldots, n$$

  - These factor scores were then used as input for another multiple linear regression model.
  - Again, the linear model assumptions were checked, residual diagnostics were examined, and the significance of each coefficient was tested.
- Both linear regression models were evaluated and compared on the test dataset using R-Squared and RMSE.

A Note On Correlation Structure

- Traditionally, one-hot encoding is applied to categorical variables so that the variables may be used appropriately in our analyses and models.
- One-hot encoding effectively creates a new dichotomous variable (1 or 0) for each category in the associated variable.
- Normally, this approach is perfectly fine before applying a machine learning model such as linear regression. However, when applying PCA or Factor Analysis, using an estimated correlation matrix, **R**, from Pearson's correlations is not technically appropriate because Pearson's correlations assume that the variables are continuous and follow a multivariate normal distribution.
- Pearson's correlation technically should not be used on dichotomous or ordinal variables (although it is often performed) [1].

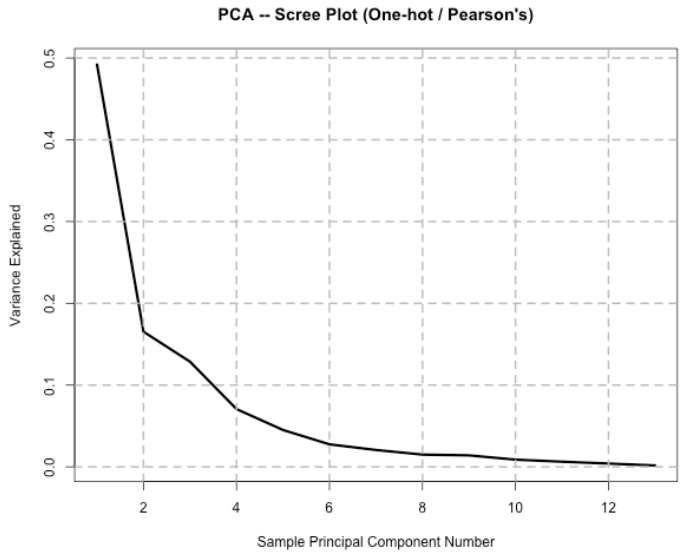
A Note On Correlation Structure

- Therefore, I also tried using a more appropriate estimate for the correlation of the categorical variables.
- I estimated the correlation matrix, **R**, using a mix of different correlation methods.
  - Numerical variables still used Pearson's correlation coefficient.
  - Dichotomous categorical variables used tetrachoric correlation [3].
  - Multi-class, ordinal, categorical variables used polychoric correlation [3] with associated natural orderings.
  - Polyserial or biserial correlations were used to estimate the elements of **R** that represented the correlation between a continuous variable and an ordinal, or dichotomous, variable.
- The analyses and modeling results from using standard one-hot encoding with Pearson's correlations were presented along with the results from using the mixed correlations.

Results & Interpretation

Principal Component Analysis (One-hot / Pearson's)

- After one-hot encoding the categorical variables, there were a total of 13 predictor variables in the training data.
- Using only Pearson's correlations, I calculated the 13 sample principal components from the training data.
- From the scree plot and the summary of variance explained, I decided to keep only the first 6 components (explained over 95% of the sample variance in training data).
- Notice that the larger weights of PC1 (≈ 0.5) are exclusively associated with variables like Engine Size, Cylinders, and Fuel Consumption levels, so PC1 may be interpreted as an average measure of engine power, or engine output.
- The relatively larger weights associated with Large Vehicle, Regular Fuel Type, and Two-Wheel Drive in PC2 indicates that PC2 may represent an aggregate measure for regular, non-premium, large passenger vehicles. For instance, this could represent a typical middle-class family-oriented car, since they are typically larger, two-wheel drive vehicles that do not require special fuel.
- While PC3 appears to be heavily dominated by the Number of Gears variable, the 3 remaining principal components may also be interpreted in a fashion as PC1 and PC2.



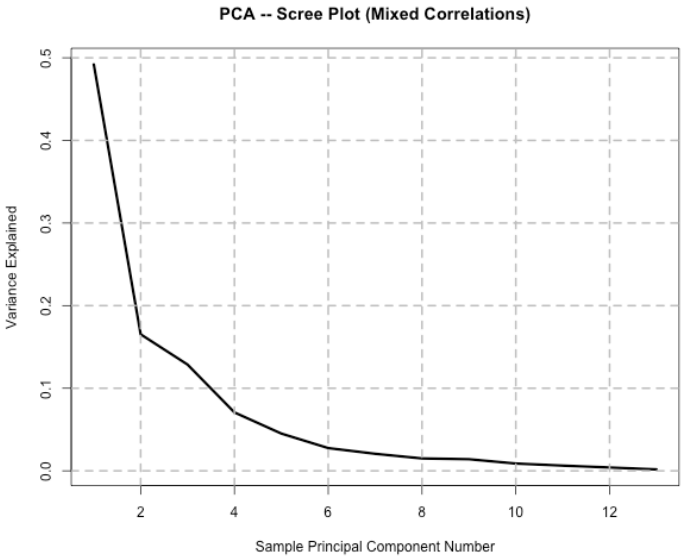
Variable Name	PC1	PC2	PC3	PC4	PC5	PC6
Vehicle.Class.FULL.SIZE	0.0141	0.0613	-0.0348	-0.0767	-0.0367	-0.0270
Vehicle.Class.LARGE.VEHICLE	0.0647	**0.3906**	-0.0180	**0.5160**	-0.1368	0.4060
Vehicle.Class.MID.SIZE	-0.0555	0.1701	-0.0662	-0.2006	0.0003	-0.1646
Engine.Size	**0.4872**	0.0641	0.0780	-0.2918	-0.3426	0.1096
Cylinders	**0.4785**	0.0466	0.0091	**-0.4433**	-0.2486	0.1764
Fuel.Type.Regular	-0.10005	**0.5293**	0.0416	0.2172	**-0.5150**	-0.0404
Fuel.Consumption.City	**0.4994**	0.0478	0.0864	0.1770	0.2791	-0.2643
Fuel.Consumption.Hwy	**0.4764**	0.0642	0.1647	**0.4044**	**0.3213**	-0.1067
Econ.Class.Luxury	0.0236	0.0183	-0.0049	-0.0512	-0.0225	-0.0502
Econ.Class.Premium	0.0070	0.2399	-0.3010	-0.1990	**0.4964**	**0.6463**
Drive.Two.Wheel.Drive	-0.0474	**0.6356**	-0.2009	-0.2761	0.2454	**-0.387**
Transmission.Type.Man	-0.0457	0.1435	-0.0090	-0.0530	0.1114	-0.2949
Number.Of.Gears	0.1872	-0.1983	**-0.9058**	0.2063	-0.1868	-0.1560

Table 1: Principal Component Weights (One-Hot Encoding / Pearson's Correlations)

Results & Interpretation

Principal Component Analysis (Mixed Correlations)

- As mentioned, I reperformed PCA using a combination of different correlation estimates depending on the variable type.
- The mixed-type correlation estimates were used to calculate the sample principal components.
- Since one-hot encoding was not performed, there were only 10 predictor variables in the training data before applying PCA.
- From the scree plot and the summary of variance explained, I decided to keep only the first 6 components (explained over 95% of the sample variance in training data).
- Again, the heavier weights on Engine Size, Cylinders, and Fuel Consumption levels indicate that PC1 may be interpreted as an average measure of engine power, or engine output.
- PC2 has a different interpretation now: The Vehicle Class and Fuel Type variables have larger positive weights while the Economic Class variable has a large negative weight, implying that PC2 may represent the difference between a vehicle's overall size and its general status.
- Note that in this case we can see that none of the components seem to be dominated by any single predictor.



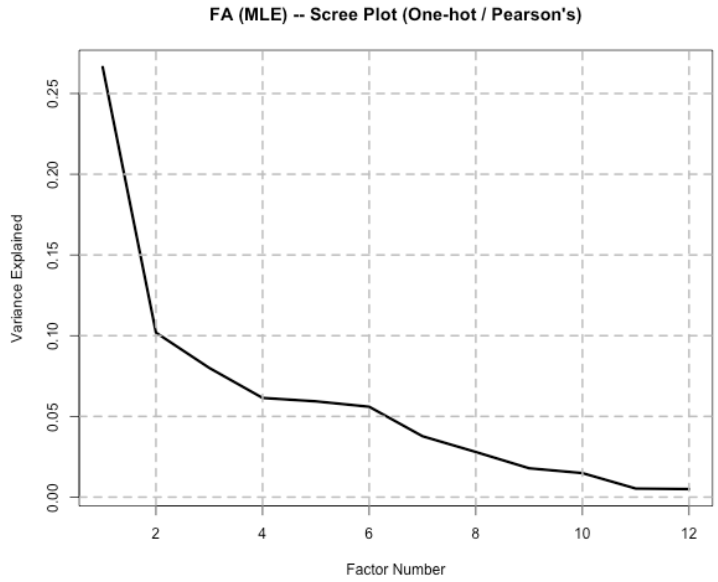
Variable Name	PC1	PC2	PC3	PC4	PC5	PC6
Vehicle.Class	-0.0086	**0.4091**	0.2681	**0.7704**	0.0454	**0.3602**
Engine.Size	**0.4407**	0.1482	-0.1941	0.0490	-0.0098	-0.2482
Cylinders	**0.4448**	0.0440	-0.1862	0.0661	-0.0705	**-0.3135**
Fuel.Type	-0.2840	**0.5183**	-0.0528	-0.1058	0.1428	**-0.3330**
Fuel.Consumption.City	**0.4548**	0.1271	-0.1917	-0.0462	0.0458	0.2235
Fuel.Consumption.Hwy	**0.4296**	0.1979	-0.1423	-0.0887	-0.0073	**0.3360**
Econ.Class	0.1638	**-0.5570**	0.2838	0.1427	**-0.3427**	0.1149
Drive	-0.1050	-0.2945	-0.5182	**0.5926**	0.0638	**-0.3552**
Transmission.Type	-0.2177	-0.1728	**-0.5670**	-0.0792	**0.3289**	**0.5292**
Number.Of.Gears	0.2247	-0.2412	**0.3546**	0.0278	**0.8606**	-0.1315

Table 2: Principal Component Weights (Mixed Correlations)

Results & Interpretation

Factor Analysis (One-hot / Pearson's)

- Used only the Pearson's correlations from the One-Hot encoded training data to compute the 13 factor loadings
- After examination of the scree plot and the variance explained by each factor, I decided to keep only the first 8 factors (explained only about 69% of the sample variance in training data).
- Interestingly, Engine Size, Cylinders, and Fuel Consumption levels all load highly on the first factor **again**, so this first factor may represent an abstract measure of engine strength, power, or output as well.
- Since premium vehicles generally require premium gasoline, this second factor seems to represent the contrast between standard vehicles and premium vehicles, with a heavy emphasis on the vehicle's fuel type.
- Many of the remaining factors appear to be dominated by a single predictor variable as highlighted in Table 3.



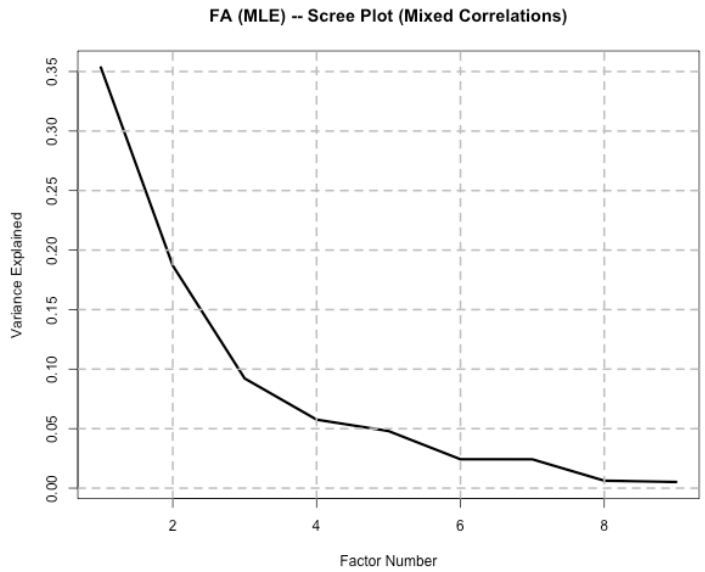
Variable Name	L1	L2	L3	L4	L5	L6	L7	L8
Vehicle.Class.FULL.SIZE	0.0604	0.0003	0.0746	**0.7405**	0.0271	0.0552	0.0016	0.0262
Vehicle.Class.LARGE.VEHICLE	0.1999	-0.2798	**0.5929**	-0.4489	-0.1454	0.4574	0.0753	-0.0423
Vehicle.Class.MID.SIZE	-0.1324	-0.0077	**-0.7625**	-0.1213	0.0043	0.0018	0.0078	-0.0494
Engine.Size	**0.9216**	-0.0018	0.0269	0.0723	0.1074	0.1385	-0.0031	0.1053
Cylinders	**0.8659**	0.1264	0.0281	0.1114	**0.3129**	0.1152	0.0362	0.1266
Fuel.Type.Regular	-0.2264	**-0.5746**	0.0181	-0.0056	-0.1189	0.0835	**0.6555**	-0.2666
Fuel.Consumption.City	**0.9293**	0.0090	0.1439	0.0291	0.0892	0.0737	-0.1032	0.1479
Fuel.Consumption.Hwy	**0.8900**	-0.0186	0.2307	-0.0869	0.0276	0.1621	-0.1410	-0.0190
Econ.Class.Luxury	0.2182	-0.0564	-0.0300	0.0385	**0.7688**	-0.0210	-0.0453	0.0493
Econ.Class.Premium	-0.0196	**0.9151**	-0.0422	0.0161	-0.0890	0.0557	-0.0864	0.1345
Crive.Two.Wheel.Drive	-0.0638	0.0481	-0.1430	0.0367	0.1178	-0.3333	-0.0420	-0.0728
Transmission.Type.Man	-0.1221	-0.0570	0.0077	-0.0578	-0.0238	**-0.5487**	-0.0066	-0.0800
Number.Of.Gears	0.1681	0.2310	0.0527	0.0394	0.0589	0.1684	-0.1079	**0.4536**

Table 3: MLE Factor Loadings (One-Hot Encoding / Pearson's Correlations)

Results & Interpretation

Factor Analysis (Mixed Correlations)

- Again, I reperformed factor analysis using a combination of different correlation estimates depending on the variable type.
- The mixed-type correlation estimates were used to calculate the factor loadings.
- Since one-hot encoding was not performed, there were only 10 predictor variables in the training data before applying factor analysis.
- After examination of the scree plot and the variance explained by each factor, I decided to keep only the first 8 factors (now explaining about 79.3% of the sample variance in training data).
- **Again,** we can see that Engine Size, Cylinders, and Fuel Consumption levels all load highly on the first factor. Therefore, the first factor may represent an abstract measure of engine strength, power, or output.
- Similarly, the second factor seems to represent the contrast between standard vehicles and premium vehicles, where fuel type is used to distinguish the two classes.
- From the larger loading values of Drive and Transmission, it appears that the third factor may represent an abstract measure of how "sporty" the vehicle is, since most pure sports cars are manual and two-wheel drive.
- The remaining factors that are not heavily dominated by a single predictor may be interpreted in a similar fashion.



Variable Name	L1	L2	L3	L4	L5	L6	L7	L8
Vehicle.Class	0.0172	-0.1780	-0.0688	**0.9727**	-0.0202	-0.1288	0.0079	-0.0017
Engine.Size	**0.9442**	0.0381	-0.0139	0.0173	0.0822	-0.1217	-0.2033	-0.0108
Cylinders	**0.9149**	0.1632	0.0411	-0.0433	0.0909	-0.1307	-0.2360	0.2189
Fuel.Type	0.2883	**-0.8914**	-0.0928	0.1442	-0.1953	-0.0457	-0.0968	0.0596
Fuel.Consumption.City	**0.9348**	0.0982	-0.0729	0.0108	0.0998	-0.0448	0.2144	-0.1735
Fuel.Consumption.Hwy	**0.8837**	0.0438	-0.1600	0.0567	0.0411	-0.0737	**0.4239**	-0.0206
Econ.Class	0.0224	**0.9614**	0.0199	-0.1056	0.1320	-0.1292	-0.0724	0.0505
Drive	0.0797	0.0790	**0.9679**	-0.0693	-0.0593	0.2040	-0.0206	0.0032
Transmission.Type	0.1903	-0.0823	0.2289	-0.1458	-0.1049	**0.9339**	-0.0030	-0.0040
Number.Of.Gears	0.1450	0.2469	-0.0641	-0.0207	**0.9504**	-0.1005	0.0031	0.0003

Table 4: MLE Factor Loadings (Mixed Correlations)

Results & Interpretation

Linear Regression Modeling

- The two sets of principal components and the two sets of factors scores were used as inputs for four different linear regression models.
- In all models, most (if not all) of the components or factors were found to be significant, and each set of components or factors demonstrated a clear linear relationship with the CO_2 emissions.
- It was found that the linear model assumptions were *approximately* satisfied in all four models, and no outliers or influential points were found in the training data.
- The testing RMSE scores shown in the table indicate excellent predictive performance from our models.
- The large testing R-Squared values tell us that each linear model using the optimal number of components, or factors, were able to explain a significant amount of the variance in the CO_2 emissions.

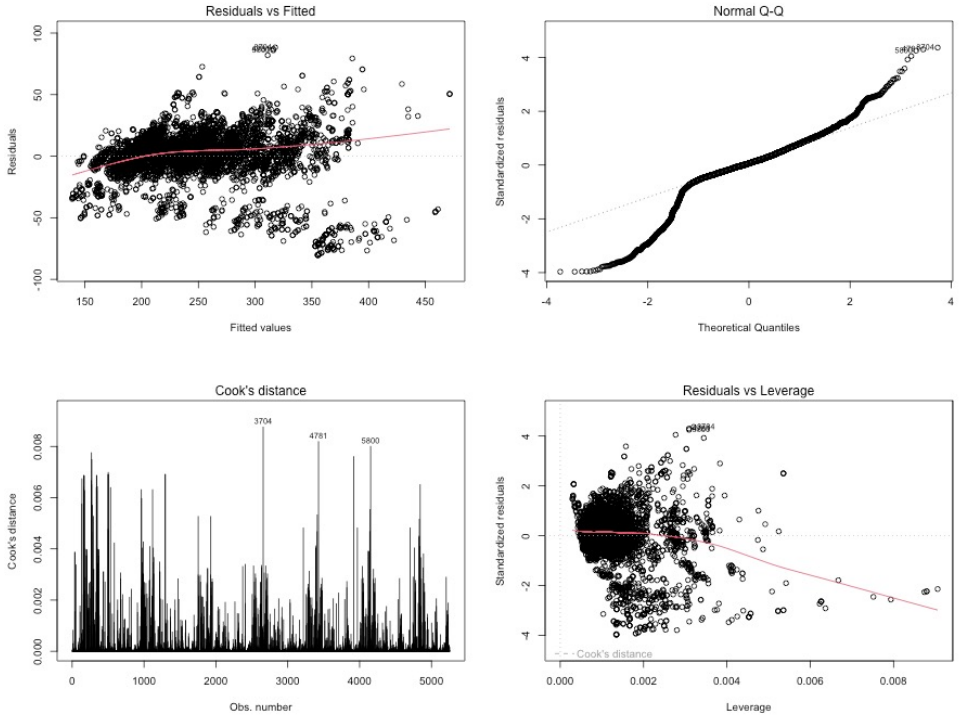


Figure: Residual Diagnostics and Outliers Plots Example (PCA w/ Regression)

Linear Model Inputs	RMSE (g/km)	R-Squared
PCA w/ One-Hot Encoding and Pearson's	19.6	0.887
PCA w/ Mixed Correlations	20.6	0.874
FA w/ One-Hot Encoding and Pearson's	19.3	0.890
FA w/ Mixed Correlations	19.3	0.891

Results & Interpretation

Notice that in each model, the coefficient of the first component, or factor, was a relatively large positive value. This implies that as the abstract measure of engine output increases (described in our interpretations of weights and loadings), we expect the CO_2 emissions to increase significantly.

	Dependent variable:			
	CO2.Emissions			
	(PCA w/ One-Hot)	(FA w/ One-Hot)	(PCA w/ Mixed)	(FA w/ Mixed)
PC1/F1	28.698***	52.115***	27.533***	51.204***
	(0.147)	(0.295)	(0.194)	(0.307)
PC2/F2	9.664***	4.947***	6.7644***	−1.226*
	(0.804)	(0.529)	(0.333)	(0.503)
PC3/F3	0.0004	14.689***	−6.328***	−12.594***
	(0.372)	(0.495)	(0.519)	(0.441)
PC4/F4	6.932***	−3.5205***	6.932***	4.866***
	(0.387)	(0.464)	(0.387)	(0.296)
PC5/F5	2.399***	0.845*	2.4714***	4.636***
	(0.489)	(0.379)	(0.418)	(0.256)
PC6/F6	3.633***	10.774***	7.429***	−2.586***
	(0.641)	(0.322)	(0.566)	(0.261)
PC7/F7		2.308***		0.966***
		(0.440)		(0.249)
PC8/F8		5.719***		2.198***
		(0.226)		(0.196)
Constant	241.767***	236.033***	266.8727***	232.776***
	(0.926)	(0.686)	(0.9161)	(1.601)
Observations	5,253	5,253	5,253	5,253
Adjusted R^2	0.883	0.888	0.873	0.889
Residual Std. Error	20.27	19.84	21.13	19.72
F Statistic	6,588***	5,187***	5,990***	5,257***
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 5: Linear Regression Coefficients for Each Model

Conclusion & Discussion

- Using PCA and Factor Analysis with different approaches to correlation estimation, I was able to successfully calculate and interpret the component weights and factor loadings.
- Although the loadings and the associated variances explained changed between each analysis, the interpretation of the first component and factor remained consistent across all four analyses.
- Linear models were successfully fit to the associated *optimal* set of components and estimated factors.
- The diagnostic tools and the testing results indicated that each model displayed a great fit and demonstrated excellent ability to predict expected CO_2 emissions.
- I found that using either PCA or factor analysis for dimensionality reduction led to similar predictive performances.
- In this case, using the polychoric and tetrachoric correlations for non-numeric variables did not significantly change the linear model's performance.

References

[1] UCLA Office of Advanced Research Computing. *How can I perform a factor analysis with categorical (or categorical and continuous) variables?* URL: https://stats.oarc.ucla.edu/stata/faq/how-can-iperform- a- factor-analysis- with- categorical- or- categorical- and- continuous- variables/. (accessed: 03.09.2023).

[2] Debajyoti Podder. *CO2 Emission by Vehicles*. URL: https://www.kaggle.com/datasets/debajyotipodder/ co2-emission-by-vehicles. (accessed: 03.09.2023).

[3] William Revelle. mixedCor: *Find correlations for mixtures of continuous, polytomous, and dichotomous variables*. URL: https : / / www.rdocumentation.org/packages/psych/versions/2.2.9/topics/mixedCor. (accessed: 03.09.2023).