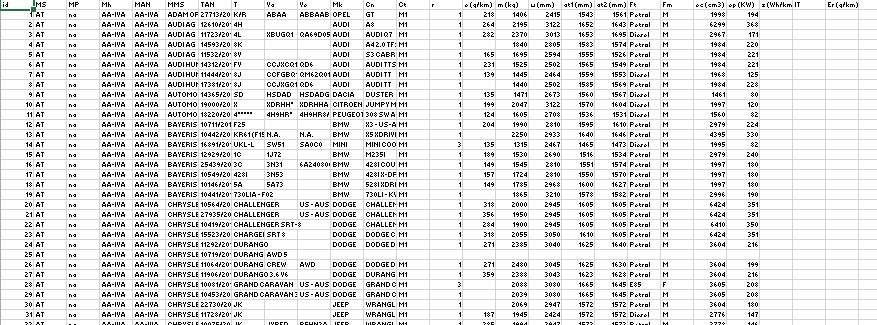
# Data Cleaning:

The first step towards analysis of any data is data cleaning.

The first step was arranging the dataset. The initial data looked like:  
  
  
After arranging the data using Excel it looked like:  
  
  
So the purpose of this arrangement was to separate the data based on the properties it possesses i.e. separating the column and placing the appropriate values to the appropriate columns.

In this carbon emission data there are many missing and unwanted values. Thus at first the number columns with missing values were handled then the columns with unwanted values was handled.  
For missing value detection a function was created (mis\_detect) and result is:  
  
 %missing

id 0.000

MS 0.000

MP 0.000

Mh 0.000

MAN 0.000

MMS 0.000

TAN 0.000

T 0.000

Va 0.011

Ve 0.000

Mk 0.000

Cn 0.000

Ct 0.000

r 5.864

e..g.km. 6.102

m..kg. 5.961

w..mm. 11.311

at1..mm. 11.781

at2..mm. 14.976

Ft 0.000

Fm 0.000

ec..cm3. 6.539

ep..KW. 26.291

z..Wh.km. 99.720

IT 0.000

Er..g.km. 99.990

Thus deleting z..wh.km. and Er..g.km. will be better as it contains information about electronic cars which is not our concern.   
To know more about the variables we checked the variable types.

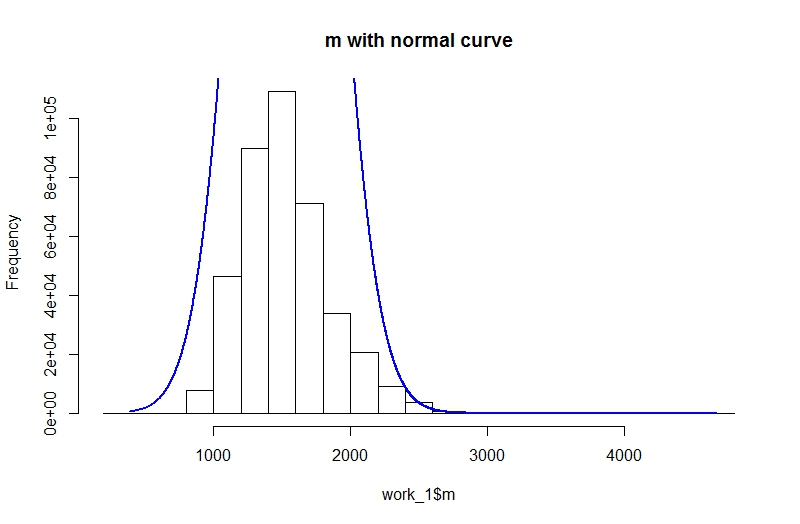
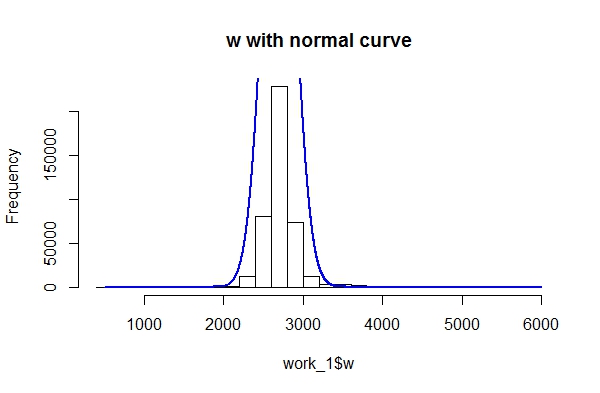
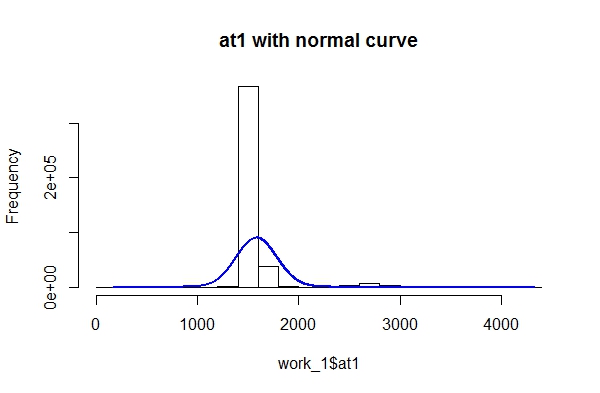
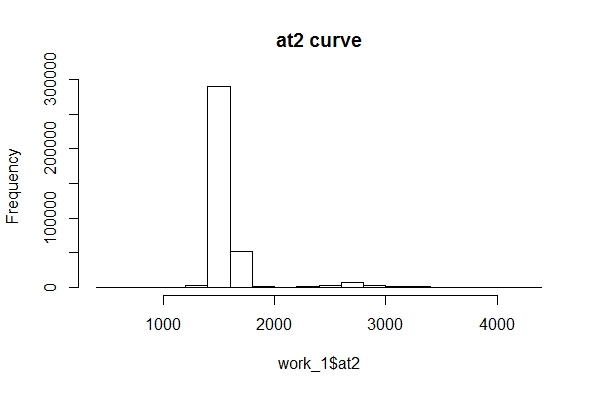
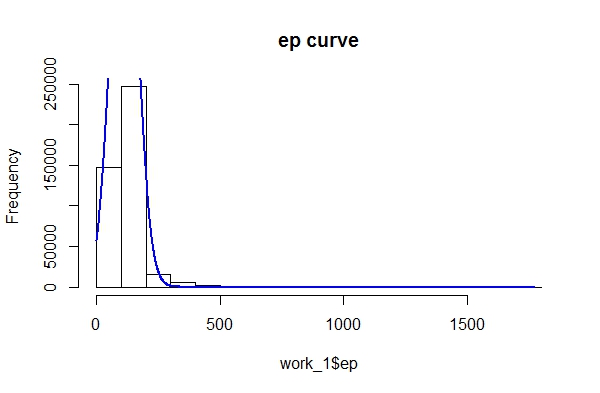
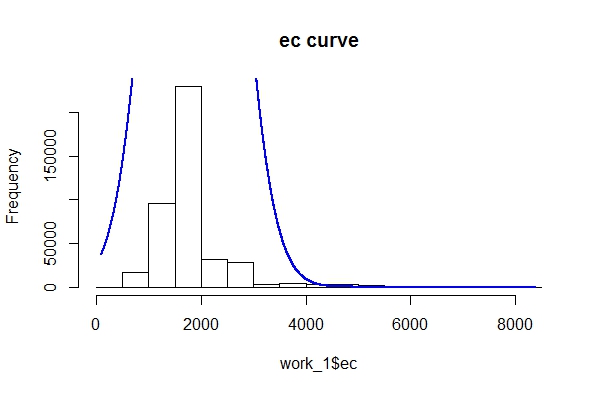
|  |
| --- |
| 'data.frame': 417939 obs. of 26 variables:  $ id : int 1 2 3 4 5 6 7 8 9 10 ...  $ MS : Factor w/ 28 levels "AT","BE","BG",..: 1 1 1 1 1 1 1 1 1 1 ...  $ MP : Factor w/ 14 levels "BMW GROUP","DAIMLER AG",..: 10 10 10 10 10 10 10 10 10 10 ...  $ Mh : Factor w/ 105 levels "","AA-IVA","AA-NSS",..: 2 2 2 2 2 2 2 2 2 2 ...  $ MAN : Factor w/ 107 levels "","AA-IVA","AA-NSS",..: 2 2 2 2 2 2 2 2 2 2 ...  $ MMS : Factor w/ 685 levels "","Å KODA","Å KODA AUTO A.S.",..: 10 39 39 39 39 41 41 41 44 48 ...  $ TAN : Factor w/ 10440 levels "","-","\*","\*\*",..: 127 58 53 87 52 84 51 100 85 107 ...  $ T : Factor w/ 3150 levels "","-","-8R","\*",..: 2120 687 692 873 886 1771 872 872 2622 2936 ...  $ Va : Factor w/ 12867 levels "","-","--","---",..: 2394 1 12077 1 1 5011 4979 5013 7416 12208 ...  $ Ve : Factor w/ 25267 levels "","'","-","--",..: 4950 1 21492 1 1 21842 22331 21842 15160 23754 ...  $ Mk : Factor w/ 266 levels ""," "," VAUXHALL",..: 196 34 34 34 34 34 34 34 73 68 ...  $ Cn : Factor w/ 14111 levels "","-"," ","#NAME?",..: 7752 2902 3327 2652 11266 3336 3334 3334 6216 8489 ...  $ Ct : Factor w/ 4 levels "","M1","M1G",..: 2 2 2 2 2 2 2 2 2 2 ...  $ r : int 1 1 1 1 1 1 1 1 1 1 ...  $ e..g.km. : int 218 264 282 NA 165 231 139 NA 135 199 ...  $ m..kg. : int 1406 2195 2370 1840 1695 1525 1445 1440 1471 2047 ...  $ w..mm. : int 2415 3122 3013 2805 2594 2502 2464 2502 2673 3122 ...  $ at1..mm. : int 1543 1652 1653 1583 1555 1565 1559 1585 1560 1570 ...  $ at2..mm. : int 1561 1643 1695 1574 1526 1549 1553 1569 1567 1604 ...  $ Ft : Factor w/ 23 levels "","Biodiesel",..: 18 18 4 18 18 18 4 18 4 4 ...  $ Fm : Factor w/ 5 levels "","B","E","F",..: 5 5 5 5 5 5 5 5 5 5 ...  $ ec..cm3. : int 1998 6299 2967 1984 1984 1984 1968 1984 1461 1997 ...  $ ep..KW. : int 194 368 171 220 221 221 125 228 80 120 ...  $ z..Wh.km.: int NA NA NA NA NA NA NA NA NA NA ...  $ IT : Factor w/ 8 levels "","715/2007/EC\*136/2014/EC (J)",..: 1 1 1 1 1 1 1 1 1 1 ...  $ Er..g.km.: num NA NA NA NA NA NA NA NA NA NA ... |
|  |
| |  | | --- | |  | |

There are some variables like MP, MMS or MH, MAN are similar kind of variables thus we can choose one of them and they would serve similar purpose asper our requirement. Based on the output required we can also remove some variables.  
Amongst the variables the continuous variables are important as the output variable is carbon emission. Thus for our analysis, based on the objective these variables are important :  
Mk(Manufacturer Co. of the car)  
MS(Member State)   
Ft(Fuel Type)  
m(Mass of the Car)  
w(Wheal Base of the Car)  
at1 & at2(Axil Base Front and Back)  
ep(engine power)  
ec(engine capacity)

# Imputation:

For imputation the 1st variable to be considered is Mk. This variable describes which company made the car.  
For this purpose at first the actual names were taken to consideration i.e. they were assembled in an excel sheet (GoogleDrive-Sheet1.csv) and used for imputation.   
So at first the ‘o’ that are wrongly inputted as ‘0’ were imputed. A function “rep0” was created (appendix II).  
Then using grep the values with 0 at the 2nd position were handled as it contained blank space in front of them.  
For removing repetition a set ‘code’ was made where the unique values were extracted and they were used for proceeding the next step analysis. The code used the distance between the letters to decide the uniqueness and thus deciding which can be the most probable one. At first a vector of codes was created, then the distance matrix and then for each row the code returned the min distance which was compared with the actual one to decide the actual word. Then the original vector coded was replaced by the original values.  
That was the job towards cleaning the data.  
In the next step all strings with only numbers as were replaced by NA so that they can be treated as missing values.  
Then unnecessary (non-alpha numeric) symbols were removed.  
With these steps the variable Mk was cleaned and now it was ready to use in analysis.  
Then the next variable i.e. ‘m’ the mass of the car was handled.  
Initially median imputation was tried for the missing values which wasn’t efficient then some huge values were removed for better analysis as they were outlier in case of this analysis.   
  
In the next step the categorical variable Ms i.e. member state was converted to lowercase ones for better analysis.  
  
In the next step the variable ‘Ft’ i.e. fuel type was cleaned.  
Where the problems like alphanumeric values were handled and blank values were converted to NA for better analysis.

# More about the Data:

For analysis initially the NA values were removed.  
Now the first thing to deal with is m, w, at1, at2, ep, ec and its relation with e as these are the variables that would be used for predicting the carbon emission based on their Fuel Type(Ft).  
So at first let’s check the variable m.  
  
The histogram shows that it can be considered normal.  
The next one is w.  
  
The distribution seems normal.  
The curve is skewed but as it is huge data thus it can be used in the model.  
  
Variable ep.  
  
Variable ec.  
  
Thus these variables can be used for model building.

The correlation between the continuous variables (without the NA values).  
 e m w at1 at2 ec ep

e 1.000 0.641 0.473 0.169 0.196 0.677 0.578

m 0.641 1.000 0.797 0.222 0.263 0.690 0.506

w 0.473 0.797 1.000 0.214 0.254 0.477 0.321

at1 0.169 0.222 0.214 1.000 0.988 0.154 0.114

at2 0.196 0.263 0.254 0.988 1.000 0.179 0.134

ec 0.677 0.690 0.477 0.154 0.179 1.000 0.772

ep 0.578 0.506 0.321 0.114 0.134 0.772 1.000

Thus we can see the correlation between the variables and there may be multicollinearity between the variables which can be handled while model’s variable selection. For variable selection at first the variables need to be normalized as all of them are in different measures.  
And the variables were very much related thus using principal components instead of the variables may be a better choice which need to be tested else AIC value should decide which variable will be in the model.

# Variable Selection for predicting carbon emission:

For variable selection we first go through PCA as we can see from the correlation table the variable have multicollinearity, after that factor scores would be used for model building as dropping a variable may cause loss of information as we have got 6 continuous variable for predicting the value of carbon emission.  
For this purpose at first the data was normalized as they have different measures and they should be at same scale for better analysis.  
Thus initially factor analysis was done and from the results we can find that at1 and at2, m and w, ep and ec are in same factor. The predictor variable is e\_un as it is normalized and yet to be renormalized to get the actual value of emission. Thus those factor scores are extracted and renamed as:  
at1 & at2 named as car\_at  
ep & ec named as car\_engine  
m & w named as car\_body  
These three variables are used model building as from the correlation matrix we can find that the multicollinearity is reduced amongst the independent variables, Bartley’s test also indicates in similar manner.

Predicting the probable value of carbon emission using factor scores and regression equation’s interpretation:   
After getting the variables suitable for predicting carbon emission value different models are thought of but multiple regression seems better model due to its simplicity and the fit will be checked by RMSE value.  
For this purpose initially the model was fitted on the entire cleaned dataset and RMSE value was checked. It was 0.48.   
For better result the dataset was divided into parts based on the fuel type labelled.  
The separate model for each fuel type showed better accuracy than the generalised one as hoped and it was confirmed by the RMSE values provided by each of the separate models.   
The results are as below:  
For generalized model the regression equation is:  
e\_un = (6.72\*e^-6) + (9.59\*e^-2) \* car\_at + (5.82\*e^-1) \* car\_engine + (4.29\*e^-1) \* car\_body  
RMSE: 0.484  
The coefficient is significant. The equeation suggests that for every increase in car\_at the e\_un increases by (9.59\*e^-2), for every increase in car\_engine the e\_un increases by (5.82^-1) and for every increase in car\_body the e\_un increases by (4.29\*e^-1).

For specific fuel types the regression models:  
Petrol:  
e\_un = 3.72 + 0.09 \* car\_at + 0.65 \* car\_engine + 0.38 \* car\_body  
RMSE: 0.31  
The coefficient is insignificant. The equeation suggests that for every increase in car\_at the e\_un increases by 0.09, for every increase in car\_engine the e\_un increases by 0.65 and for every increase in car\_body the e\_un increases by 0.38.  
  
Diesel:  
e\_un = -0.33 + 0.10 \* car\_at + 0.44 \* car\_engine + 0.68 \* car\_body  
RMSE: 0.26  
.The equeation suggests that for every increase in car\_at the e\_un increases by 0.10, for every increase in car\_engine the e\_un increases by 0.44 and for every increase in car\_body the e\_un increases by 0.68.

Biodiesel:  
e\_un = 1.37 + 7.28 \* car\_at – 1.96 \* car\_engine – 0.88 \* car\_body  
RMSE: 0.20  
.The equeation suggests that for every increase in car\_at the e\_un increases by 0.7.28, for every increase in car\_engine the e\_un decreases by 1.96 and for every increase in car\_body the e\_un decreases by 0.88.

diesel electric:  
insufficient data

e85:

e\_un = 0.54 + 0.13 \* car\_at + 0.64 \* car\_engine + 0.76 \* car\_body  
RMSE: 0.109  
The equeation suggests that for every increase in car\_at the e\_un increases by 0.13, for every increase in car\_engine the e\_un increases by 0.64 and for every increase in car\_body the e\_un increases by 0.76.

hydrogen:  
Only two observartions thus prediction is unreliable

lpg:

e\_un = 0.53 + 0.14 \* car\_at + 0.80 \* car\_engine + 0.39 \* car\_body  
RMSE: 0.29  
The equeation suggests that for every increase in car\_at the e\_un increases by 0.14, for every increase in car\_engine the e\_un increases by 0.80 and for every increase in car\_body the e\_un increases by 0.39.

ng biomethane:

e\_un = -0.33 + 0.06 \* car\_at + 0.42 \* car\_engine + 0.69 \* car\_body  
RMSE: 0.18  
The equeation suggests that for every increase in car\_at the e\_un increases by 0.06, for every increase in car\_engine the e\_un increases by 0.42 and for every increase in car\_body the e\_un increases by 0.69.

petrol electric

Insufficient data for model building.

petrol gas:

Insufficient data for model building.

# Conclusion:

Thus from the above observations and models we can conclude that the individual models are better than generalized models as they display less RMSE value than the generalized models. For the insufficient data cases generalized models can be used for prediction.