# Data Analysis for Individual Carbon Footprints

Question - What are the predictors of individual carbon emissions, and how can we estimate personal carbon footprints based on these predictors?

Group B

Boyang Zhao, Fangyi Tian, Hesen Huang, Yapeng Guo, Yiyan Meng

## Abstract

**Background**

Climate change has emerged as one of the most pressing challenges facing humanity in the 21st century. Reducing greenhouse gas emissions, particularly carbon dioxide, is crucial to mitigating its adverse effects. In addition to industrial and governmental actions, the daily lifestyles of individuals are also closely related to carbon emissions and thus affect climate change. Therefore, we are interested in the impact of various individual lifestyle factors on carbon emissions and conduct comprehensive research.

**Objectives**

Our study aims to explore and quantify the relationship between Individual lifestyle and carbon emissions, identify the predictors of individual carbon emissions and develop predictive models, which could provide strategies for improving personal lifestyle, thereby reducing greenhouse gas emissions and protecting the environment.

**Study Design and Findings**

For the study purposes mentioned above, our study design encompassed a rigorous Exploratory Data Analysis to convert some data types and identify and visualize key data patterns, followed by correlation analysis to determine the strength of relationships between variables and select our independent variables (numeric). We selected 'carbon.emissions' as the dependent variable Y, and selected 12 attributes as our independent variables after EDA (10 of them are numeric, 2 are category).

Next, we conducted hypothesis tests for two category independent variables using the Wilcoxon Rank-Sum Test. The analysis of variance (ANOVA) was used to assess whether there were differences in carbon emissions across the population with different body types. We also considered the effects of 'body.type' and 'sex' on carbon emissions by using a two-way ANOVA.

Then, we constructed a multiple linear regression model to quantify the impact of these variables on carbon emissions and compared it with a Lasso regression model. Finally, we chose the Lasso model with 12 independent variables and ensured model robustness through model evaluation. It is a fairly good model prediction and explanation capabilities, with an R-squared value of 0.7486. The key individual factors affected carbon emissions are mainly travel and transportation factors, personal waste factors, body type and gender are the factors that have the most significant impact on it.

All the above analyses were performed using R software version 4.3.1 and RStudio version 2023.06.2+561.

## 1 Introduction

The increase in greenhouse gases in the atmosphere is disrupting the environment and causing severe climate change (Pandey et al., 2010). The consequences of climate change, such as ecosystem damage, pose significant risks to human society and the natural world (VijayaVenkataRaman et al., 2012). To effectively mitigate climate change, we should call for reductions in carbon dioxide emissions from all sources, including individual households.

As an important indicator related to environmental protection, carbon footprint measures the total amount of greenhouse gas emissions related to individuals, organizations, or products (Wiedmann&Minx, 2008). Previous research focused more on the impact of government and industry on the environment but lacked attention to individual-level carbon emissions. Therefore, our study aims to explore the impact of the multifaceted nature of individual life on carbon emissions.

We applied comprehensive statistical analysis methods to synthetic datasets to quantify the relationship between the lifestyle of individuals and their carbon footprint and develop predictive models that provide actionable insights and strategies for people to enhance their environmental awareness, which promotes them to participate in fighting against climate change and protect the environment.

## 2 Data Analysis

### 2.1 Data Source and Collection

Our dataset was sourced from Kaggle website, an open-access platform (https://www.kaggle.com/datasets/dumanmesut/individual-carbon-footprint-calculation), ensuring transparency and reproducibility of our results. These data are generated and calculated based on a combination of weights from various studies and sites currently used to calculate carbon emissions, which are very close to real-world values for an individual's carbon footprint and its associated lifestyle factors. The dataset was uploaded by 5 contributors Mesut Duman, Ecem Bayindir, Burhan Yildiz and Huseyin Baytar in February 2024.

### 2.2 Data Description

The dataset consists of 10,000 samples and 20 features, encompassing a wide range of lifestyle characteristics. We looked at the characteristics of all variables and selected ‘CarbonEmission’ as the dependent variable which represents an individual's total carbon emissions. The other 19 variables are ‘Body.Type’, ‘Sex’, ‘Diet’, ‘How.Often.Shower’, ‘Heating.Energy.Source’, ‘Transport’, ‘Vehicle.Type’, ‘Social.Activity’, ‘Monthly.Grocery.Bill’, ‘Frequency.of.Traveling.by.Air’, ‘Vehicle.Monthly.Distance.Km’, ‘Waste.Bag.Size’, ‘Waste.Bag.Weekly.Count’, ‘How.Long.TV.PC.Daily.Hour’, ‘How.Many.New.Clothes.Monthly’, ‘How.Long.Internet.Daily.Hour’, ‘Energy.efficiency’, ‘Recycling’ and ‘Cooking_With’, which captured various aspects of an individual's lifestyle that potentially influence their carbon footprint (Table 1). We looked at their features in next step EDA and selected 12 renamed independent variables and converted some types of them for analysis (Table 3).

**Table 1. Original Variable**

| Variable | Type | Variable | Type |
|---|---|---|---|
| CarbonEmission | numeric | Heating.Energy.Source | category |
| Body.Type | category | Transport | category |
| Sex | category | Vehicle.Type | category |
| Diet | category | Social.Activity | category |

| | | | |
|---|---|---|---|
| How.Often.Shower | category | Monthly.Grocery.Bill | numeric |
| Frequency.of.Traveling.by.Air | category | Vehicle.Monthly.Distance.Km | numeric |
| Waste.Bag.Size | category | How.Long.Internet.Daily.Hour | numeric |
| Waste.Bag.Weekly.Count | numeric | Energy.efficiency | category |
| How.Long.TV.PC.Daily.Hour | numeric | Recycling | category |
| How.Many.New.Clothes.Monthly | numeric | Cooking_With | category |

### 2.3 Exploratory Data Analysis, EDA

### 2.3.1 Data Preprocessing

After browsing the general structure and content of the data, we first renamed the variable names to improve readability and ease of use. Then we checked the missing values and dealt with them. Next we converted 5 categorical variables which had meaningful ordinal relationships, body.type, social.act, airtvl.freq, wastebag.size, and energy.eff, to numeric values based on their ordinal levels. The remaining categorical variables were converted to factors. Finally, We checked outliers for numerical variables using the interquartile range (IQR) method and decided not to remove them after visualization (Figure 1).

### 2.3.2 Visualization

1. Distribution of Carbon Emission (Y): We created a histogram plot for our dependent variable carbon.emission to view its distribution (Figure 2).

2. Analysis of 19 Independent Variables

We analyze our numerical variables and categorical respectively to identify key factors that significantly influence carbon emissions and select them for our model.

For 7 category independent variables: We used box plots to see their approximate relationship with the dependent variable carbon.emission. Then we dropped 5 variables and remained 2 variables (sex and transport) based on the analysis of the box plots (Figure 3). We further converted transport to a binary variable and displayed its relationship with carbon.emission (Figure 4). For 12 numerical independent variables: We used histograms and scatter plots to understand their distributions and potential relationships with carbon emissions. And all numerical variables showed some degree of relationship with carbon emissions from the scatter plot. We showed one variable veh.distance in (Figure 5).

To further explore the relationship between numerical variables and dependent variables, we roughly looked at their correlation.

### 2.3.3 Correlation

We calculated the correlation matrix using the Pearson method, which is the standard approach for measuring the linear relationship between variables (Table 2). Then we displayed it and used a circular correlation plot to visualize the coefficients, in which the sizes and colors of the circles denote the relationships (Figure 6). Then we dropped one variable tvpc.hour based on the correlation.

After we completed our EDA, and finally chose 12 variables as our independent variables, including 2 category variables: sex and transport; 10 numeric variables body.type, energy.eff, internet.hour, social.act, groc.bill, wastebag.size, wastebag.num, new.clothes, airtvl.freq, and veh.distance. Our dependent variable is carbon.emission (Table 3).

### 2.4 Hypothesis Test

we performed two hypothesis tests respectively to investigate whether there are significant differences in carbon emissions based on two key factors: sex and transportation preferences. To select the appropriate statistical test, we first assessed the normality of the carbon emission data for each group using histograms and the Shapiro-Wilk normality tests (Table 4). Based on the non-normality of the results we chose the non-parametric tests Wilcoxon rank-sum tests for both. Then we formulated the null and alternative hypotheses.

For sex, we have $H_0$: $\text{median}_{\text{male}} = \text{median}_{\text{female}}$ vs. $H_A$: $\text{median}_{\text{male}} \neq \text{median}_{\text{female}}$

For transport, we have H0: $\text{median}_{\text{environ}} = \text{median}_{\text{unenviron}}$ vs. $H_A$: $\text{median}_{\text{environ}} \neq \text{median}_{\text{unenviron}}$

And we conducted the Wilcoxon rank-sum tests in R specifying the carbon emission as the dependent variable and the grouping factor (sex or transportation) as the independent variable. with a significance level of 0.05 and rejected both null hypotheses based on the very small p-value (Figure 7).

### 2.5 Analysis of Variance, ANOVA

we conducted both one-way ANOVA for body.type and two-way ANOVA for investigating the effects of body type and sex on carbon emissions.

We first performed the one-way ANOVA to assess the impact of body type on carbon emissions. The body type variable (body.type) was converted as a categorical factor for ANOVA with four levels: underweight, normal, overweight, and obese. The carbon emission (carbon.emission) was the dependent variable. To further explore the differences between specific groups, we also performed post-hoc tests Tukey's Honest Significant Difference (HSD) test. And finally, we drew diagnostic plots.

Next, we conducted a two-way ANOVA in R to examine the main effects of body type (body.type) and sex on carbon emissions as well as their interaction effect (see APPENDIX). After that we assessed the significance of the main effects and interaction effect and visualized the interaction between body type and sex (Figure 11). The TukeyHSD() was also used to perform post-hoc comparisons for the main effects and interaction effect (see APPENDIX).

### 2.6 Multiple Linear Regression

As Table 3, we finally chose 12 variables as our independent variables and carbon.emission as our dependent variables. Then we specified a linear regression model with carbon.emssion as our dependent variable. The model included all selected predictors and was fit using the ordinary least squares (OLS) method. Then we used these variables to fit the multiple linear regression model. Next, we calculated Variance inflation factors (VIFs) to assess multicollinearity among the independent variables. Generally, a VIF greater than 5 or 10 indicates high correlation and potential redundancy among predictor variables. We did not choose any interaction items and then we used summary() to see the output (Figure 12).

### 2.7 Model Selection and Evaluation

We began by dividing our dataset into two equal-sized samples: the in-sample (training set) and the out-sample (test set), which allows us to train and evaluate the models on separate data, ensuring an unbiased assessment of model performance. We first applied the Lasso (Least Absolute Shrinkage and Selection Operator) regularization technique to the training set. To determine the optimal value of the regularization parameter (lambda), we performed cross-validation and visualized the results (Figure 13).

Then we got the coefficients of our Lasso model (Figure 14) and compared the performance of the Lasso model to our multiple linear regression model. We fitted both models to the training set and evaluated their performance on the test set. For each model, we calculated the MSE and the coefficient of determination (R-squared) by hand (Table 5). Then based on the results we chose Lasso model and did some residual visualization analysis (Figure 15). Finally, we got our Lasso model:

***carbon.emission = -1278.22 + 180.95 × body.type + 95.13 × social.act + 0.89 × groc.bill + 436.68 × airtvl.freq + 0.20 × veh.distance + 135.67 × wastebag.size + 79.09 × wastebag.num + 13.87 × new.clothes + 5.59 × internet.hour - 25.77 × energy.eff + 364.03 × sex.num + 124.22 × trans.num***

## 3. Result

### 3.1 Exploratory Data Analysis, EDA

#### 3.1.1 Data Preprocessing

After we renamed all the variables in this step, we checked the missing values. We found that although the dataset did not contain any explicit missing values, the 'vehicle.type' variable had blank values. Considering that many people may not drive or have cars, meaning that they do not use fuel. So we carefully decided to replace these values instead of deleting them.

Next we converted some types of our variables, and used IQR method to check outliers for numerical variables. And we found that there were 1294 outliers in ’veh.distance’ and 346 in ‘carbon.emission’. So we visualized it by box plot (Figure 1).

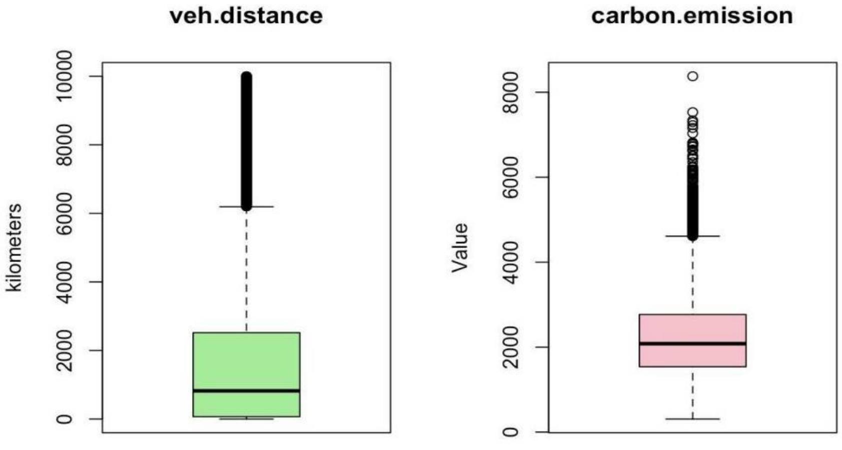


**Figure 1. Outliers in Two Numeric Variables**

We thought that the number of outliers for veh.distance and carbon.emission are not small, so removing them may affect the analysis. And they are not necessarily errors. We could infer that a very high vehicle distance might cause traveling frequently and high energy consumption from the plot, thus related to high carbon.emission. So there is no need to delete them because they might reveal interesting patterns.

#### 3.1.2 Visualization

1. Distribution of Carbon Emission (Y)

The histogram of carbon.emission showed that its distribution was right-skewed, with most values concentrated in the lower range and some higher extreme values, meaning that the mean of individual carbon emission was greater than the median of individual carbon emission (Figure 2).

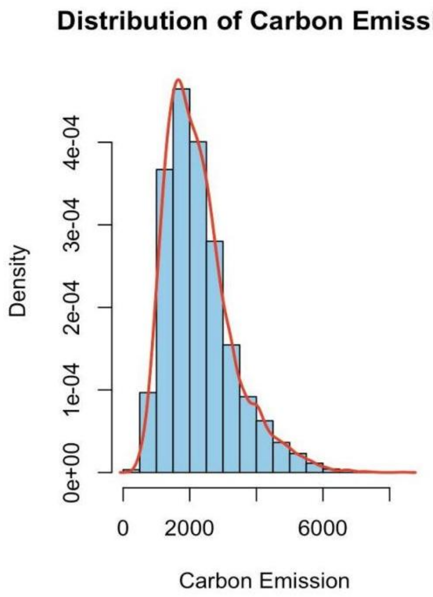


**Figure 2. Distribution of Carbon Emissions**

2. Analysis of Independent Variables

(1) For 7 category independent variables: By viewing the boxplots, we removed variables with too many categories (veh.type, recycling, cooking) and those showing tiny differences between groups (diet, shower.freq, heat.src) and kept two key variables sex and transport with clear relationships with carbon emissions to simplify the model (Figure 3).

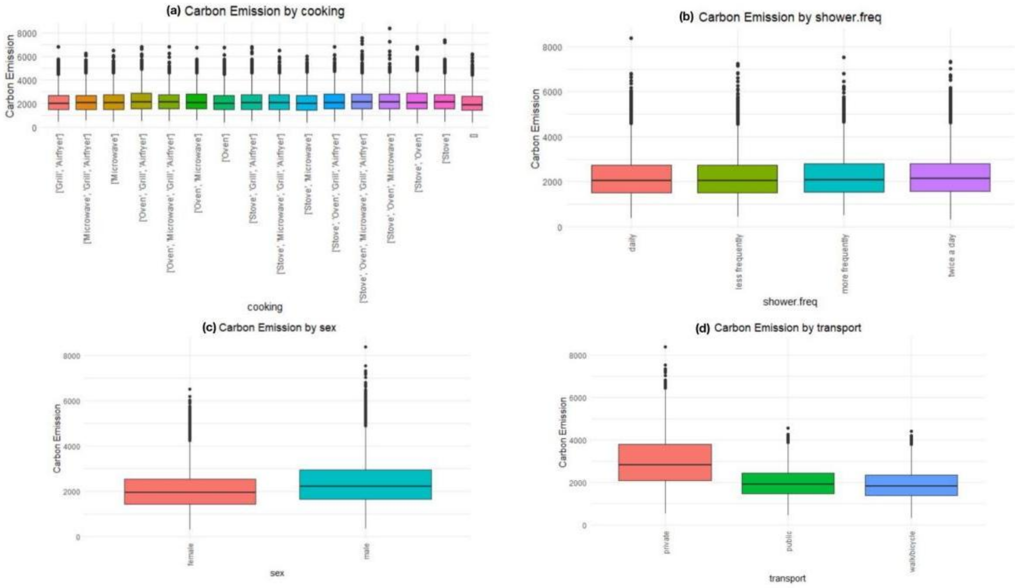


**Figure 3. Categorical variables vs. carbon emissions combined box plot**

Figure 3.(a) taking the variable cooking as an example shows too many categories; Figure 3.(b) taking the variable shower.freq as an example, shows very small differences between groups; (c) and (d) are the two variables we selected, sex and transport with dependent variable, showing a clear relationship

Specifically, we further analyzed the variable transport, categorizing it into 'environ' (environmentally friendly mode containing 'public' and 'walk/bicycle') and 'unenviron' (not environmentally friendly containing 'private'). Then draw a new boxplot plot to show its relationship with Y after being converted (Figure 4). Although the number of samples in the two groups in the variable transport is very different, "unenviron" has 3279 samples and "environ" has 6721 samples. However, we still include it in the analysis because exploring transportation modes is a key factor affecting carbon emissions. And the box plot revealed a significant difference in carbon emissions between the two groups in variable transport, with 'unenviron' showing higher median emissions.

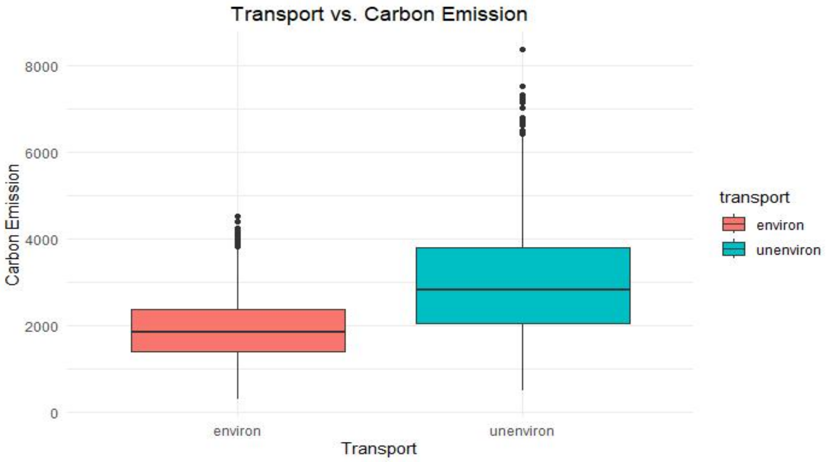


**Figure 4. Modified Transport vs. Carbon Emission Boxplot**

(2) For 12 numeric independent variables:

To display the distribution of each variable, we used histogram (see APPENDIX). From histograms, most have no obvious characteristics, except for the distribution of veh.distance, which was right-skewed. And the scatter plots were used to scatter understand the potential relationships between the independent variables with carbon emissions. Most of them showed some degree of relationship with carbon emissions from the scatter plots (see APPENDIX). Notably, one variable veh.distance exhibited a pretty clear positive correlation so we displayed it here specially (Figure 5).

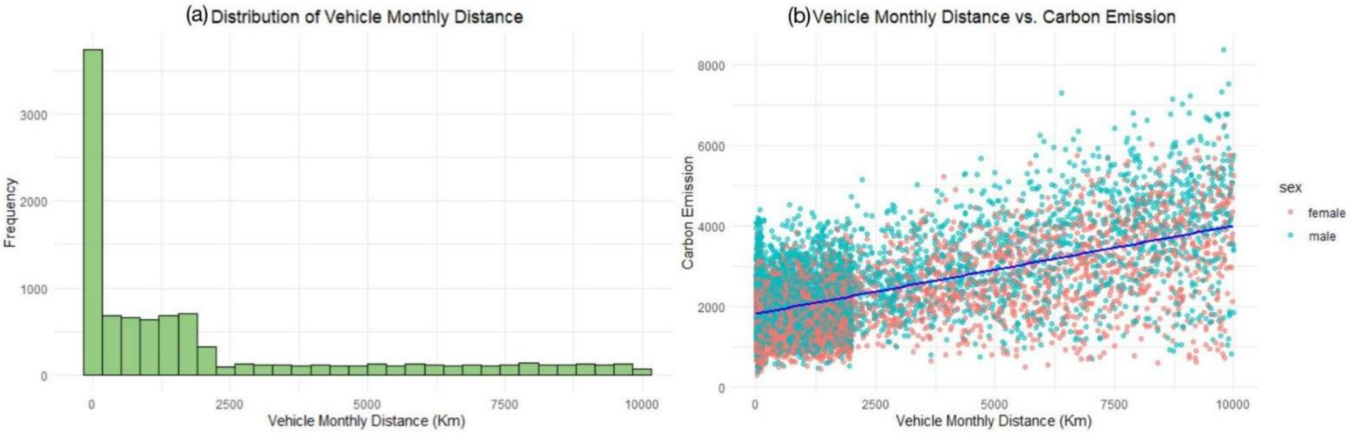


**Figure 5. Variable veh.distance Related Graph**

Figure 5.(a) The histogram shows the distribution of veh.distance, which is skewed to the right; Figure 5.(b) shows the relationship between veh.distance and depent variable carbon emission with a positive correlation.

### 3.1.3 Correlation

The correlation analysis revealed several notable findings regarding the relationships between the variables and carbon emissions (Table 2).

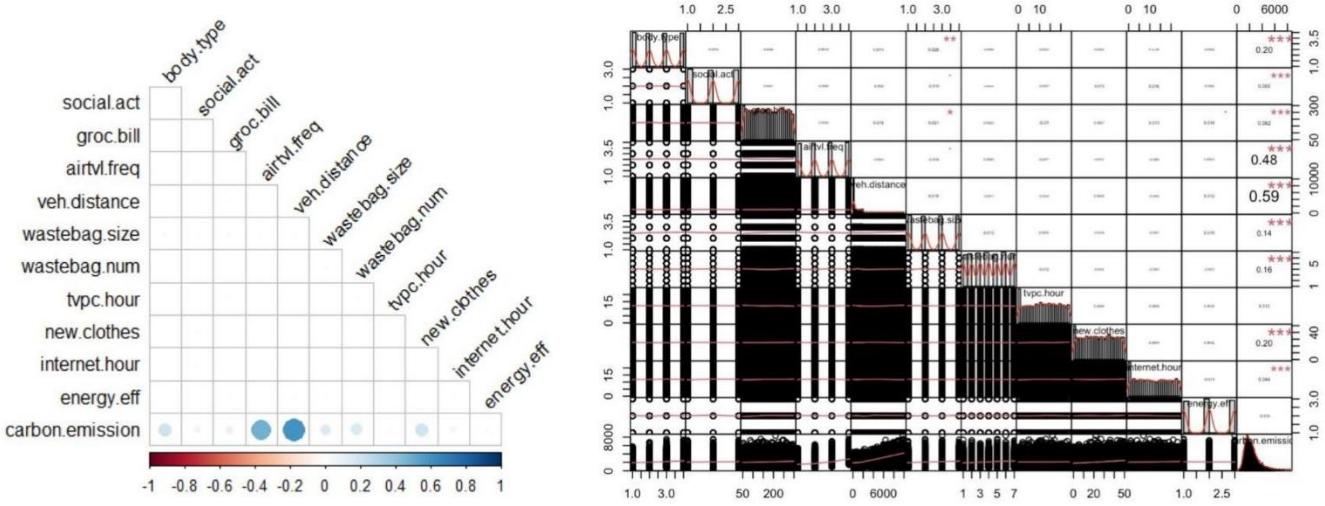
**Table 2. Correlation of Numeric Variables with Carbon Emission**

| Variable | Correlation with carbon.emission | Variable | Correlation with carbon.emission |
|---|---|---|---|
| energy.eff | -0.01371188 | wastebag.num | 0.15919337 |
| tvpc.hour | 0.01298500 | new.clothes | 0.19888735 |
| internet.hour | 0.04387803 | body.type | 0.20316917 |
| social.act | 0.05537568 | airtvl.freq | 0.47848675 |
| groc.bill | 0.08158658 | veh.distance | 0.59417130 |
| wastebag.size | 0.14239526 | carbon.emission | 1.00000000 |

Among the independent variables, vehicle distance (veh.distance) exhibited the strongest positive correlation with carbon emissions, with a correlation coefficient of 0.5942. This suggests that individuals who travel greater distances by vehicle tend to have higher carbon emissions. Air travel frequency (airtvl.freq) also showed a moderately strong positive correlation with carbon emissions (0.4785), indicating that frequent air travel contributes to higher carbon footprints.

Other variables that demonstrated positive correlations with carbon emissions included body type (0.2032), new clothes purchases (0.1989), waste bag size (0.1424), and waste bag number (0.1592). These correlations suggest that factors such as higher body weight, increased consumption of new clothes, larger waste bag sizes, and more frequent waste disposal are associated with higher carbon emissions. In addition, TV & PC Hour (tvpc.hour) has littlecorrelation with carbon.emission (0.01298500), so we decided to delete it in our model.

The correlation plot provided a clear visual representation of the pairwise correlations, highlighting the strength and direction of the associations (Figure 6).



(a) Circular Correlation Plot

(b) Correlation Matrix with Plots

Figure 6. Correlation between Numerical Variables and Dependent Variable

Figure 6.(a) is circular correlation plot; Figure 6.(b) displays the correlation matrix along with histograms and scatter plots

In fact, we could see from the figure that there is almost no correlation between the independent variables, which also greatly reduces the impact of multicollinearity.

Up to this point, we have selected all the independent variables that we are interested in and suitable for modeling, as shown in Table 3.

**Table 3. Variable Characteristics (N = 10000)**

| Variable | Type | Range |
|---|---|---|
| sex (X) | category | 1 (female) / 2 (male) |
| transport (X) | category | 1 (eniviron) /2 (uneniviron) |
| body.type (X) | numeric | 1-4 |
| social.act (X) | numeric | 1-3 |
| groc.bill (X) | numeric | 50-299 |
| airtvl.freq (X) | numeric | 1-4 |
| veh.distance (X) | numeric | 0-9999 |
| watebag.num (X) | numeric | 1-7 |
| watebag.size (X) | numeric | 1-4 |
| new.clothes (X) | numeric | 0-50 |
| internet.hour (X) | numeric | 0-24 |

| | | |
|---|---|---|
| energy.eff (X) | numeric | 1-3 |
| carbon.emission (Y) | numeric | 306-8377 |

### 3.2 Hypothesis Test

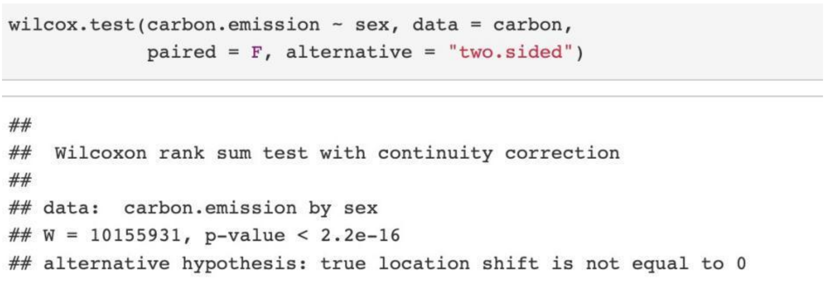
We conducted two hypothesis tests, respectively for sex and transport.

First, to select the appropriate statistical test, we used histograms (see APPENDIX) and the Shapiro-Willk normality test. The histograms revealed that the data did not follow a normal distribution for both sex and transportation groups. The Shapiro-Wilk test further confirmed this, with p-values less than 0.05, indicating a significant deviation from normality (Table 4).

**Table 4. Shapiro-Wilk Test Statistic for Normality Check of Carbon Emission Data**

| Variable | Shapiro-Wilk Test Statistic (W) | p-value |
|---|---|---|
| Male | 0.92683 | < 2.2e-16 |
| Female | 0.93526 | < 2.2e-16 |
| Environ | 0.98214 | < 2.2e-16 |
| Unenviron | 0.97658 | < 2.2e-16 |

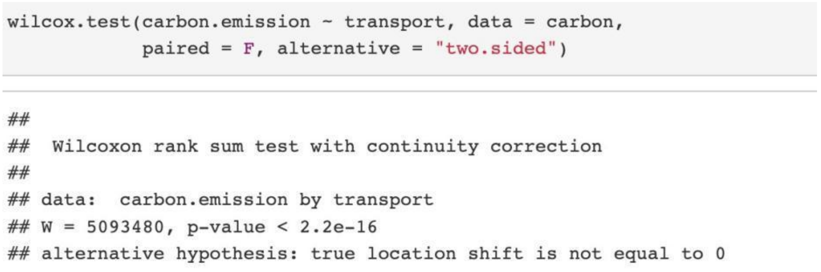
Given the non-normality of the data, we opted for the Wilcoxon rank-sum test and made our hypothesis. Then we conducted the Wilcoxon rank-sum test and got the results (Figure 7).



```
wilcox.test(carbon.emission ~ sex, data = carbon,
            paired = F, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  carbon.emission by sex
## W = 10155931, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

**Figure 7.(a) Wilcoxon Rank-Sum Test Results for sex**



```
wilcox.test(carbon.emission ~ transport, data = carbon,
            paired = F, alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  carbon.emission by transport
## W = 5093480, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

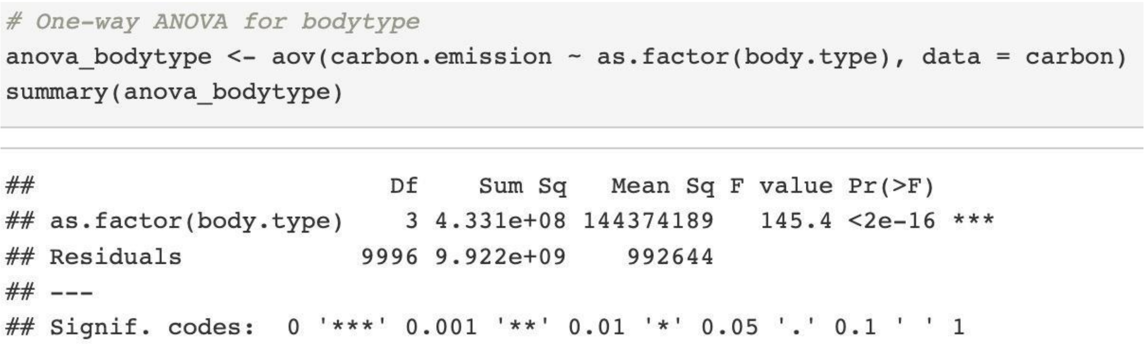
**Figure 7.(b)Wilcoxon Rank-Sum Test Results for transport**

The Wilcoxon rank-sum test for sex yielded a p-value less than 2.2e-16, which is well below the chosen significance level of 0.05. We can reject the null hypothesis and concluded that there was a statistically significant difference in median carbon emissions between males and females. Males have a higher median carbon emission (2235) compared to females (1942). Similarly, the Wilcoxon rank-sum test for transportation preferences also resulted in a p-value less than 2.2e-16, indicating a significant difference in median carbon emissions between environmentally friendly and non-environmentally friendly modes of transportation. The non-environmentally friendly transportation has a considerably higher median carbon emission (2817) compared to environmentally friendly transportation (1841).

## 3.3 ANOVA

### 3.3.1 One-way ANOVA

The results showed a highly significant effect of body type on carbon emissions with p < 2e-16 (Figure 8).



```
# One-way ANOVA for bodytype
anova_bodytype <- aov(carbon.emission ~ as.factor(body.type), data = carbon)
summary(anova_bodytype)
```

```
##                        Df    Sum Sq   Mean Sq F value Pr(>F)    
## as.factor(body.type)    3 4.331e+08 144374189   145.4 <2e-16 ***
## Residuals            9996 9.922e+09    992644                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Figure 8. R Output of One-way ANOVA**

At 0.05 level of significance, the results of the one-way ANOVA revealed a significant effect of body type on carbon emissions (F = 145.4, p < 2e-16). The p-value was much smaller than the significance level of 0.05, indicating strong evidence against the null hypothesis. So we can conclude that all body type groups do not have all the same mean carbon emissions, meaning that at least one mean carbon emission value differs significantly from the others.

Then the post-hoc test (Tukey's HSD test) revealed significant differences in mean carbon emissions between all pairs of body type groups (all p-values < 0.001). This suggests that each body type group has a distinct mean carbon emission (See APPENDIX).

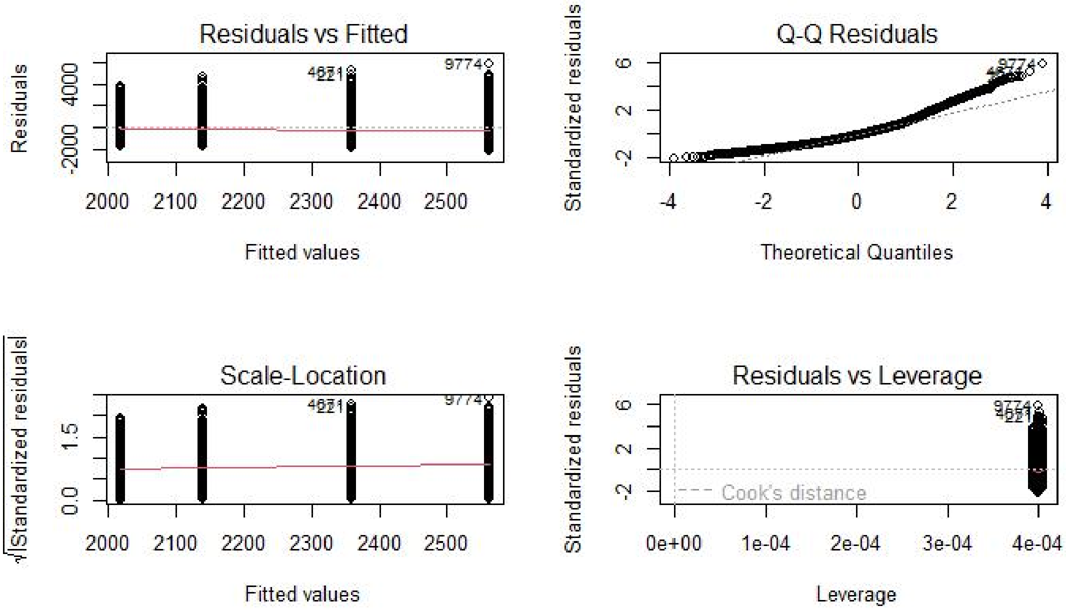


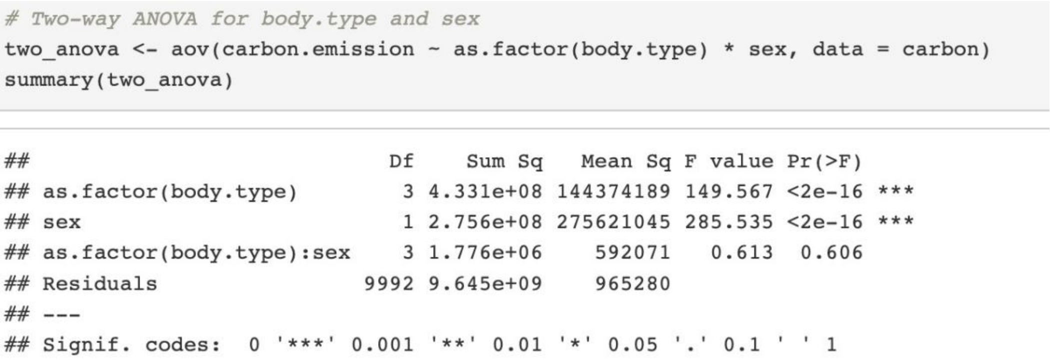
**Figure 9. Diagnostic Plots for One-way ANOVA**

From the diagnostic plot, we cannot see that the residual distribution has a clear pattern (Residuals vs. Fitted Plot and Scale-Location Plot), which means that the residuals may have homoskedasticity. A normal

Q-Q Plot shows that most of the points fall on the 45-degree line, but there is a deviation in the tail, so the residuals are not necessarily normally distributed and require further inspection (Figure 9).

### 3.3.2 Two-way ANOVA

The two-way ANOVA results indicated significant main effects of both body type (p < 2e-16) and sex (p < 2e-16) on carbon emissions. However, the interaction effect between body type and sex was not significant (p = 0.606) at 0.05 level of significance (Figure 10).



```
# Two-way ANOVA for body.type and sex
two_anova <- aov(carbon.emission ~ as.factor(body.type) * sex, data = carbon)
summary(two_anova)
```

```
##                              Df    Sum Sq   Mean Sq F value Pr(>F)    
## as.factor(body.type)          3 4.331e+08 144374189 149.567 <2e-16 ***
## sex                           1 2.756e+08 275621045 285.535 <2e-16 ***
## as.factor(body.type):sex      3 1.776e+06    592071   0.613  0.606    
## Residuals                  9992 9.645e+09    965280                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Figure 10. R Output of Two-way ANOVA**

Our interaction plot showed the mean carbon emissions for each sex within each body type category. The parallel lines suggest that the trends in carbon emissions across body types are similar for both males and females (Figure 11). And we also did a post-hoc test for this two-way ANOVA, which shows the similar results: there are significant differences in carbon emissions between all body types, and, a significant difference was found between males and females. But the interaction between body type and sex was not (see APPENDIX).

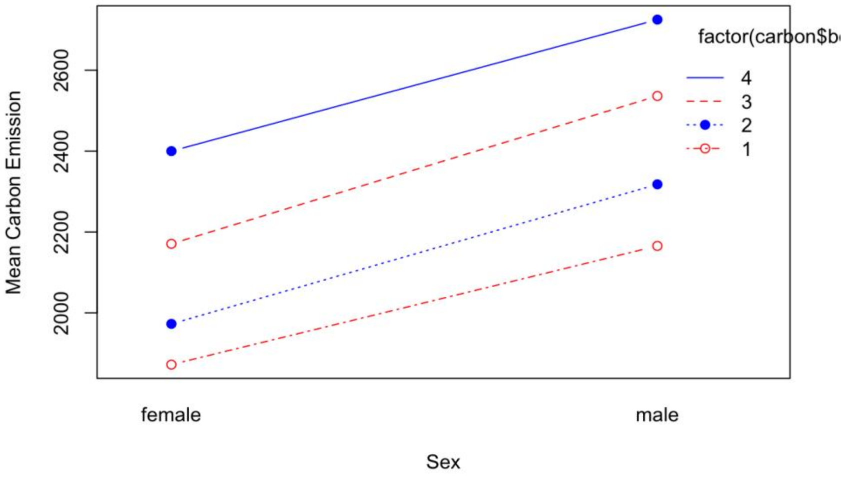
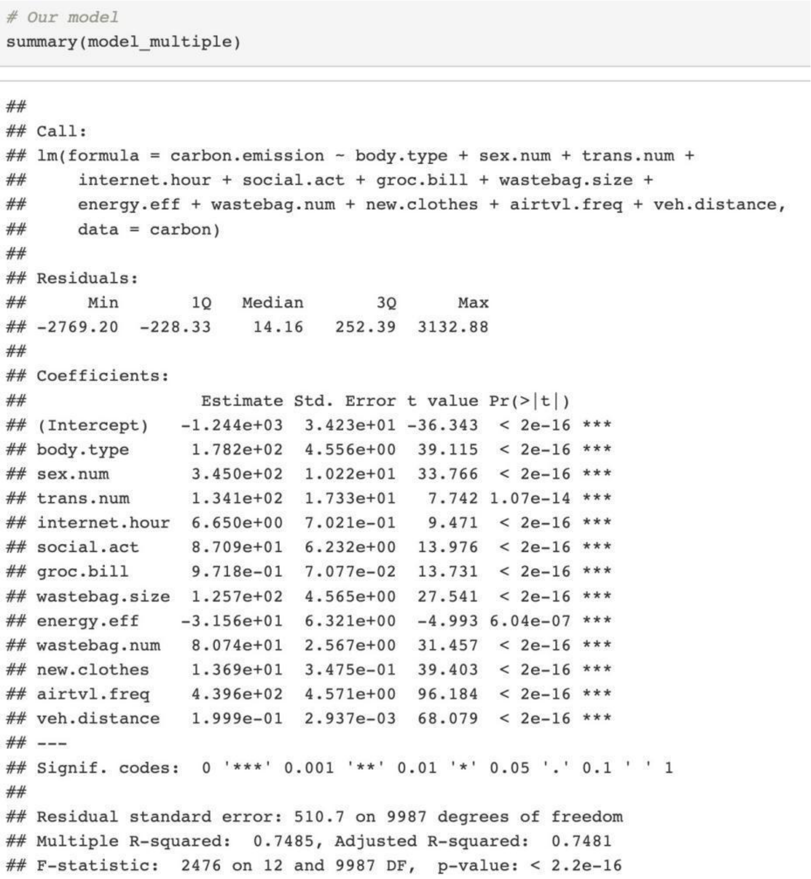


**Figure 11. Body Type and Sex Interaction Plot for Mean Carbon Emission**

### 3.4 Multiple Linear Regression

We first got our VIFs (see APPENDIX), all of our VIF values are close to 1, suggesting no significant multicollinearity among the predictors. Then we got the output of our model (Figure 12).



```
# Our model
summary(model_multiple)
```

```
##
## Call:
## lm(formula = carbon.emission ~ body.type + sex.num + trans.num +
##     internet.hour + social.act + groc.bill + wastebag.size +
##     energy.eff + wastebag.num + new.clothes + airtvl.freq + veh.distance,
##     data = carbon)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2769.20  -228.33    14.16   252.39  3132.88
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.244e+03  3.423e+01 -36.343  < 2e-16 ***
## body.type      1.782e+02  4.556e+00  39.115  < 2e-16 ***
## sex.num        3.450e+02  1.022e+01  33.766  < 2e-16 ***
## trans.num      1.341e+02  1.733e+01   7.742 1.07e-14 ***
## internet.hour  6.650e+00  7.021e-01   9.471  < 2e-16 ***
## social.act     8.709e+01  6.232e+00  13.976  < 2e-16 ***
## groc.bill      9.718e-01  7.077e-02  13.731  < 2e-16 ***
## wastebag.size  1.257e+02  4.565e+00  27.541  < 2e-16 ***
## energy.eff    -3.156e+01  6.321e+00  -4.993 6.04e-07 ***
## wastebag.num   8.074e+01  2.567e+00  31.457  < 2e-16 ***
## new.clothes    1.369e+01  3.475e-01  39.403  < 2e-16 ***
## airtvl.freq    4.396e+02  4.571e+00  96.184  < 2e-16 ***
## veh.distance   1.999e-01  2.937e-03  68.079  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 510.7 on 9987 degrees of freedom
## Multiple R-squared:  0.7485, Adjusted R-squared:  0.7481
## F-statistic:  2476 on 12 and 9987 DF,  p-value: < 2.2e-16
```

**Figure 12. Multiple Linear Regression Model**

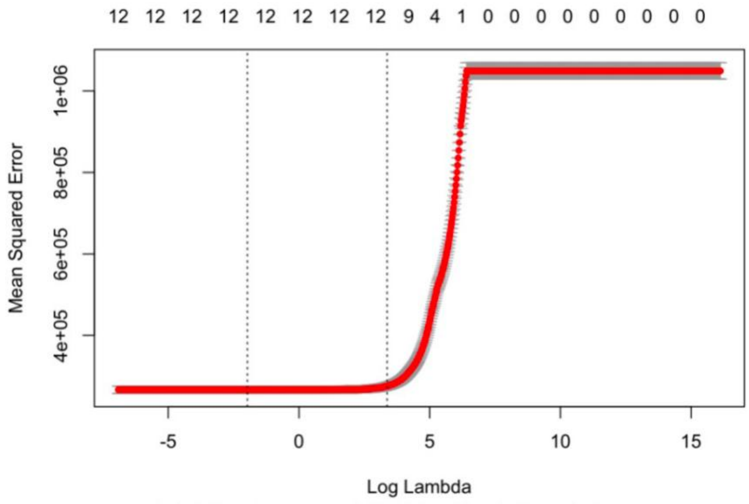
From the output, we can see the Multiple R-squared is 0.7485, meaning that our fitted multiple linear regression model explained approximately 74.85% of the variance in carbon emissions. This is a substantial amount. Besides, all variables are statistically significant (p < 0.05). The estimated coefficients for each independent variable indicate their impact on carbon emissions. For example, a one-unit increase in new.clothes is associated with a 13.69 increase in carbon emissions, holding other variables constant. And the F-statistic for the overall model was 2476 with a p-value less than 2.2e-16, confirming that the model was statistically significant.

The multiple linear regression model can be presented in equation form as follows:

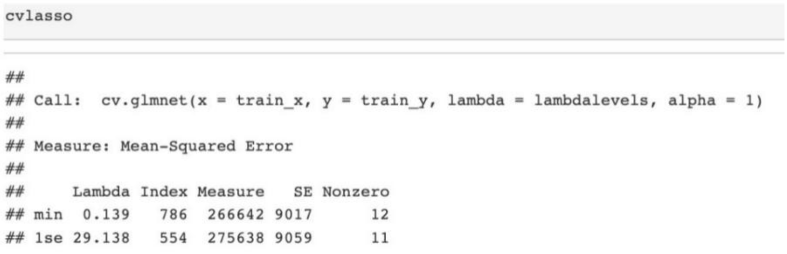
Carbon Emission=-1244+178.2×Body Type+345×Sex+134.1×Transport Type+6.65×Internet Hours+87.09×Social Activities+0.9718×Grocery Bills+125.7×Waste Bag Size-31.56×Energy Efficiency+80.74×Waste Bag Number+13.69×New Clothes+439.6×Air Travel Frequency+0.1999×Vehicle Distance

### 3.5 Model Selection and Evaluation

The dotted vertical line on the left shows the lambda that gets the best MSE.  The cross-validation results for the Lasso model showed that the best-fitting lambda value was 0.139, corresponding to a model with 12 non-zero coefficients (Figure 13).



(a). The Average MSE for Each Lambda



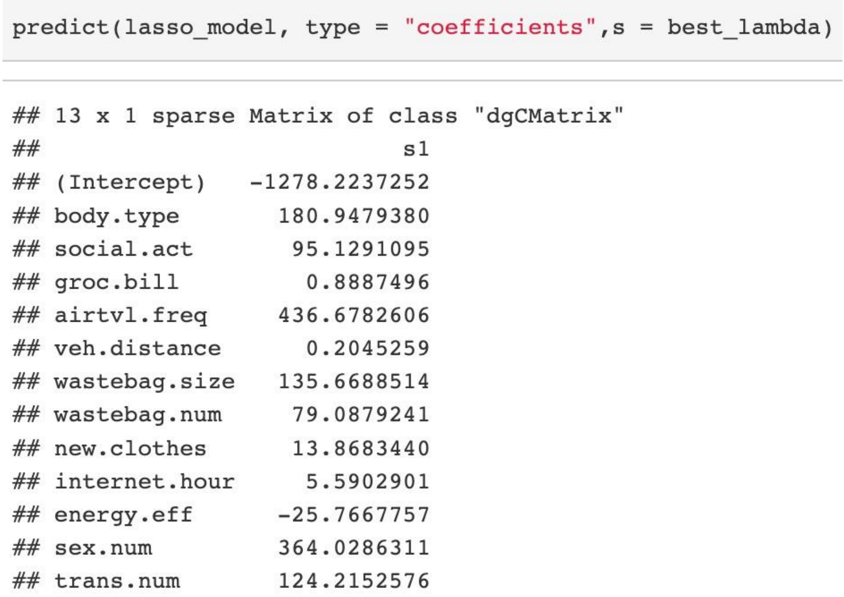
```
cvlasso
```

```
##
## Call:  cv.glmnet(x = train_x, y = train_y, lambda = lambdalevels, alpha = 1)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure   SE Nonzero
## min  0.139   786  266642 9017      12
## 1se 29.138   554  275638 9059      11
```

(b) Cross-validation Results

**Figure13. MSE and Lambda in Cross-Validation**

And then we can got our coefficients of Lasso model in R using predict() function (Figure 14).



```
predict(lasso_model, type = "coefficients",s = best_lambda)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                          s1
## (Intercept)   -1278.2237252
## body.type       180.9479380
## social.act       95.1291095
## groc.bill         0.8887496
## airtvl.freq     436.6782606
## veh.distance      0.2045259
## wastebag.size   135.6688514
## wastebag.num     79.0879241
## new.clothes      13.8683440
## internet.hour     5.5902901
## energy.eff      -25.7667757
## sex.num         364.0286311
## trans.num       124.2152576
```

**Figure 14. Coefficients of Lasso Model**

The comparison of the Lasso model and the multiple linear regression model revealed that the Lasso model slightly outperformed the regression model in terms of both MSE and R-squared. The Lasso model had an MSE of 256900 and an R-squared of 0.7486306, while the regression model had an MSE of 256910.5 and an R-squared of 0.7486205 (Table 5). Although the differences were small, the Lasso model demonstrated better predictive accuracy. So we decided to choose the Lasso Model.

**Table 5. Comparison of Stepwise Regression and Lasso Regression Models**

| **Model** | **MSE** | **R-squared** |
|---|---|---|
| Multiple Linear Regression | 256910.5 | 0.7486205 |
| Lasso Regression | 256900 | 0.7486306 |

Therefore, our model is:

***carbon.emission = -1278.22 + 180.95 × body.type + 95.13 × social.act + 0.89 × groc.bill + 436.68 × airtvl.freq + 0.20 × veh.distance + 135.67 × wastebag.size + 79.09 × wastebag.num + 13.87 × new.clothes + 5.59 × internet.hour - 25.77 × energy.eff + 364.03 × sex.num + 124.22 × trans.num***

Now we got the model and its R-squared and MSE. We tried to do some residual visualization analysis (Figure 15).

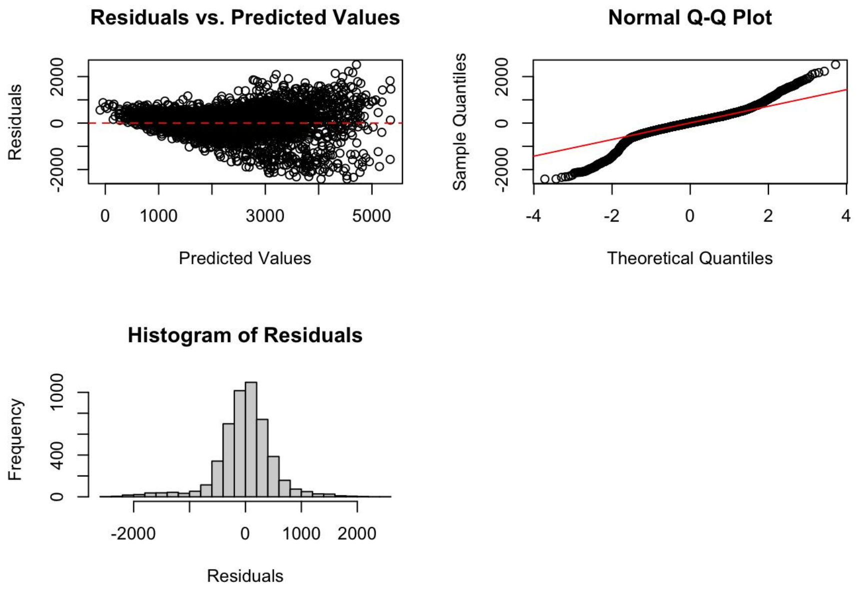


**Figure 15. Residual Analysis for Lasso Model**

The residual analysis of the Lasso model showed no significant patterns or deviations from the assumptions of linearity and homoscedasticity. The residuals were randomly scattered around the horizontal line at zero in the residuals vs. predicted values plot, indicating a good fit of the model. The Q-Q plot shows that the residuals are distributed roughly in a straight line along 45 degrees, with only the tails deviating slightly, indicating that the residuals are roughly normally distributed. The histogram of the residuals also exhibited a roughly symmetric distribution centered around zero.

Based on these results, we selected the Lasso model as our final model for predicting individual carbon emissions. The Lasso model effectively identified the most important predictors while providing good predictive accuracy. The final Lasso model included the following 12 Independent variables: sex, transport, body.type, energy.eff, internet.hour, social.act, groc.bill, wastebag.size, wastebag.num, new.clothes, airtvl.freq, and veh.distance.

## 4.Discussions

Our study aimed to investigate the factors influencing individual carbon emissions using a dataset sourced from Kaggle. Through exploratory data analysis (EDA), hypothesis testing, and regression modeling, we gained insights into the relationships between various lifestyle factors and carbon emissions. Our Lasso model was selected because of its good model explanation, fitting and prediction capabilities, which had a great R-squared 0.7486306 with 12 coefficients. These 12 variables were selected to predict carbon emission, sex, transport, body.type, energy.eff, internet.hour, social.act, groc.bill, wastebag.size, wastebag.num, new.clothes, airtvl.freq, and veh.distance, which had significant coefficients.

Besides, combining correlation analysis and the Lasso regression model, we believed that the key factors affecting personal carbon emission among the 12 independent variables were travel factors (including distance and frequency, veh.distance and airtvl.freq), gender factor (sex), and personal garbage factor (including the size and amount of garbage, wastebag.size and wastebag.num) and body type (body.type).

The insights gained from this study have important implications for promoting sustainable lifestyles and mitigating the environmental impact of human activities. By understanding the key drivers of individual carbon emissions, policymakers, environmental organizations, and individuals themselves can develop targeted interventions and educational campaigns to encourage eco-friendly behaviors and reduce carbon footprints. For example, there should be some policy Interventions. Policymakers should consider promoting sustainable transportation options and energy-efficient practices to reduce carbon emissions at the individual level. Promote calls for waste classification and disposal. Environmental protection public education campaigns can also be carried out to raise people's awareness.

However, our study has some the limitations.

Data Limitation: Firstly, the dataset used in this study was synthetically generated, which raises concerns about its representativeness and generalizability to real-world scenarios. Synthetic data may not capture the full complexity and variability of real-life carbon footprint data. So in the future we could validate the findings using real-world data.

In addition, when discussing the transport variable, we found that after conversion to a binary variable, the "environ" group had 6721 samples, while the "unenviron" group had only 3279 samples. This imbalance in sample size may affect the reliability and accuracy of analytical results. The power of statistical tests may be reduced. In future studies, data resampling techniques may be considered to balance the sample size between groups, or sensitivity analysis to imbalance may be performed. Despite this limitation, our analysis still provides valuable insights into the relationship between transport variables and carbon emissions. There are significant differences in carbon emissions between environmentally friendly and non-environmentally friendly transport modes, underscoring the importance of encouraging sustainable transport choices.

Importance of independent variable units: When interpreting the coefficients in the regression model, it is crucial to consider the scales and units of the independent variables. For instance, the strong positive correlation between vehicle distance (veh.distance) and carbon emissions in the correlation analysis may not be directly reflected in the small coefficient value in the regression model. This discrepancy can be attributed to the different scales of the variables. Standardizing or normalizing the variables before fitting the model could help mitigate the impact of variable scales and facilitate more intuitive interpretation of the coefficients.

Develop different models: We could explore various models. For example, another avenue for future exploration is the transformation of the dependent variable, carbon emission, into categorical levels (e.g., low, medium, high) and the development of a logistic regression model. This approach could provide insights into the factors that distinguish individuals with different levels of carbon emissions. Or we could explore non-linear relationships and employ more advanced machine learning algorithms.

Nonetheless, this study serves as a valuable starting point for further investigations into the determinants of individual carbon footprints. By continuing to refine our understanding of the factors influencing carbon emissions, we can develop more effective strategies for promoting sustainable living and combating climate change.

## 5.Conclusions

Our research finally succeeded in developing a suitable personal carbon emission prediction model via a series of statistical analysis methods using 12 independent variables. We believed that travel and transportation factors, personal waste factors, body type and gender are the factors that have the most significant impact on it.

We can realize from this analysis that individuals have a significant impact on the environment and greenhouse gas emissions. Therefore, solving the problem of climate change also requires individual efforts. Our model can help individuals predict and assess their carbon impact, provide a reference for achieving a low-carbon lifestyle, and develop effective emission reduction plans to protect the beautiful environment of our home planet.

## APPENDIX

See attached files:

1. Carbon Emission.csv
 (https://www.kaggle.com/datasets/dumanmesut/individual-carbon-footprint-calculation/data)
2. 7343G2Final.pdf

**Contribution of each member:**

Fangyi Tian: Did most of the project. Including the main data analysis and the main report. Statistical analysis - Designed and conducted the project, and wrote the code. Conducted EDA, especially data cleansing, correlation analysis, and variable selection. Conducted hypothesis test and ANOVA. Developed the multiple linear regression model and Lasso model, then compared, selected, and evaluated the model. Report - Organized, wrote, and revised the project report.

Yiyan Meng: Participating in the selection and analysis of dataset. Review the project report and the related code, then extract and organize the content in the project report. According to the organized content, design layout and production of ppt content.

Yapeng Guo: Finding datasets on the internet, conducting initial missing value imputation and variable processing, participating in hypothesis testing and ANOVA result output, developing code for variable analysis and visualization. Writing the draft report.

Hesen Huang: Basic data description-get the summary of the data set, helped to understand the type of each features and the data structure, did different types of data visualization, correlation of some features with carbon emission and ANOVA, prior multiple linear regression with numerical features or other features.

Boyang Zhao: Analyze the regression models, using the MSE and R-squared method and doing some residuals analysis and getting the plots, and evaluate the model by combining methods from the perspective of linearity, homoscedasticity and explanatory power.

## References

Pandey, D., Agrawal, M., & Pandey, J. S. (2010). Carbon footprint: current methods of estimation. *Environmental Monitoring and Assessment (Print)*, *178*(1–4), 135–160. https://doi.org/10.1007/s10661-010-1678-y

VijayaVenkataRaman, S., Iniyan, S., & Goić, R. (2012). A review of climate change, mitigation and adaptation. *Renewable & Sustainable Energy Reviews*, *16*(1), 878–897. https://doi.org/10.1016/j.rser.2011.09.009

Wiedmann, T., & Minx, J. (2008). A definition of 'carbon footprint'. *Ecological economics research trends*, *1*(2008), 1-11.