A Short Monograph on Analysis of Variance (ANOVA)

TO SERVE AS A REFRESHER FOR PGP-DSBA

Index

Contents

List of Figures

List of Tables

1. Importance of Variance in Analytics

1.1 Why understanding variability in data is important?

Advancement of technologies has made it easy for all organizations to collect and store a vast amount of data. Each data point is different from the others, and understanding, analyzing and generating actionable insight from the data has become essential for the organizations to remain ahead of their competition. Understanding data is to a large extent synonymous to understanding the various sources of variability in the data.

Variability is an inherent property of an attribute (random variable). Take for example the “height” of a person. The property that not all adult human beings are of the same height, is due to variability in the height distribution. To predict the height of a randomly selected adult human being with any degree of accuracy, it is important to understand which factors, if any, are responsible for the difference in height.

Source of variability may be of two primary types: Systematic and Error (random or chance).

When one or more factors can be identified contributing to the variability, it is known as a Systematic Source of variation. One systematic source of variation of height is gender. Typically, (on the average) male adult human beings are taller than female adult human beings.

When the variability cannot be attributed to any known source, it is known as error. If there is a height difference between two adult twin siblings of the same sex, the difference is purely due to chance or error.

Any analytical or prediction problem tries to identify as many systematic sources of variation as possible. The higher proportion of variation can be attributed to systematic sources, the more accurate the predictions will be.

There are many statistical techniques to identify the systematic sources of variation in a data set. Analysis of Variance (ANOVA) is one of the simplest techniques that identify one or more factors that may contribute to the source of variability.

2. What is Analysis of Variance?

2.1 Definition of Analysis of Variance (ANOVA)

The formal definition of Analysis of variance (ANOVA): ANOVA is a statistical technique that assumes that the observed response is coming from more than one population and tests the hypothesis that at least one population mean is different from the rest.

The basic concept of ANOVA is to separate the total variability in a dataset into two types, the variability that can be attributed to specified causes and the variation that can be attributed to chance or error.

The objective of the ANOVA: Analysis **o**f **Va**riance (ANOVA) is a hypothesis testing technique that is used to determine whether the means of more than two populations are identical. The underlying assumption is that the heterogeneity or variability in the data is due to the fact that the data is coming from more than two different normal populations whose variance is the same.

This technique is used in various problems such as in comparing yields of the crop from several varieties of seeds, the gasoline mileage of various types of automobiles, satisfaction score of customers with respect to mobile network services in different locations, etc. This technique has application in various fields such as sociology, economics, marketing, laboratory experiments, etc.

A few important definitions related to ANOVA are given below:

Experimental design is the plan used to collect the data. The basic purpose of setting an experiment is to observe the impact of one or more factors on the observed variable.

The ***factor*** is an independent explanatory variable with several levels. Each level of the factor represents a different population.

The ***response is the Dependent variable*** which is continuous and assumed to follow a normal distribution

Consider, an example where interest lies in comparing the weekly volume of sales by different teams of sales executives. Here, the sales team is the factor with multiple levels and weekly sales volume is the response. It is conjectured that weekly volume will depend on the team. One point needs to be clarified here. Since ANOVA is not applicable for comparison of two population means (two-sample problem is handled through a t-statistic whereas ANOVA employs an F-statistic), in this monograph we will always assume that the factor has more than two levels.

Types of ANOVA: We have discussed two types of Analysis of Variance problems in detail:

I. **One-way ANOVA:** When the response depends on a single factor
II. **Two-way ANOVA:** When the response depends on two factors who may or may not interact between themselves

Case Study:

Traffic management inspector in a certain city wants to understand whether carbon emissions from different cars are different. The inspector has reasons to believe that Fuel type (LPG, Petrol or Petrol (E85-Flex Fuel)) and car manufacturer (Audi, BMW, Ford, Volvo) may be the factors responsible for differences in carbon emission. For this purpose, she has taken random samples from all registered cars on the road in that city and would like to compare the amount of carbon emission release due to fuel type and/or manufacturers.

This problem is essentially a problem of identification of the source(s) of variation in the data. ANOVA will be applied to see whether

- Carbon emission depends on fuel type only (One-way ANOVA)
- Carbon emission depends on manufacturer only (One-way ANOVA)
- Carbon emission depends on both fuel type and manufacturer both (Two-way ANOVA)

3. One- way ANOVA

3.1 One-way Analysis of Variance

One-way ANOVA tests the null hypothesis

$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_c$

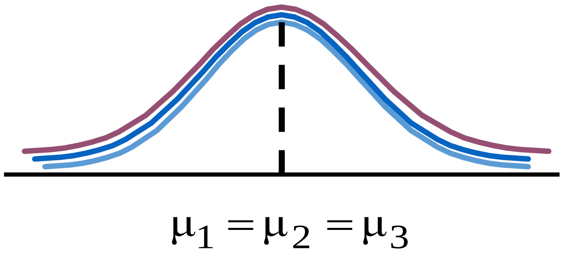

Figure 1: Null Hypothesis of One-Way ANOVA
(Courtesy: Malhotra, N., Hall, J., Shaw, M., & Oppenheim, P. (2006). Marketing research: An applied orientation. Pearson Education Australia.)

Against the alternative

H_a: At least one population mean is different from the rest.

This is another way of saying that all population means are not identical. Note that, this is NOT the same as saying all population means are different.

Figure 1 illustrates three normal populations whose means are identical. Figure 2 below illustrates two different cases. The LHS figure demonstrates that means of population 1 and 2 are identical whereas the mean of population 3 is different from them. The RHS figure demonstrates that all three population means are different.

Note that it is assumed that the population variances are all equal. If population variances are different ANOVA cannot be applied. Assumptions for ANOVA are discussed below.

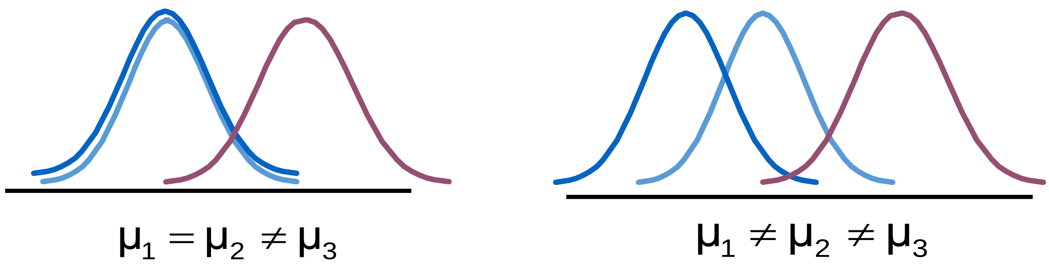

Figure 2: Alternative Hypothesis of One-way ANOVA
(Courtesy: Malhotra, N., Hall, J., Shaw, M., & Oppenheim, P. (2006). Marketing research: An applied orientation. Pearson Education Australia.)

Assumptions for ANOVA

1. All populations under consideration have normal distribution
2. All populations under consideration have equal variances.
3. The sample is a random sample, i.e. the observations are collected independently of each other.

Formal tests exist to test Assumptions 1 and 2. Assumption 3 is ensured through the sampling mechanism.

Before the case-study is taken up, let us consider the motivation and rationale behind ANOVA.

3.2 How variation is partitioned into two parts in One-way ANOVA?

Given a set of n observations from the same population $Y_1, Y_2, \ldots, Y_n$, their variance can be represented as: $var(Y) = \frac{\sum(Y-\bar{Y})^2}{n-1}$, where the numerator is the sum of squared deviations from mean ($\bar{Y}$) and the denominator ($n-1$) is the corresponding degrees of freedom. We will call the numerator as Total Sum of Squares (SST).

Consider now the set of observations coming from c populations, $j = 1, 2, \ldots c$, where n_j observations are coming from the j^{th} population. $\sum_{j=1}^{c} n_j = n$, sample size. Total sum of squares (SST) for this data set may be expressed as:

$$\text{SST} = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \bar{\bar{Y}})^2$$

Where $\bar{\bar{Y}} = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n_j} Y_{ij}$ is the overall mean.

The total sum of squares can be divided into two additive and independent components

- variation among groups (SSB) and,
- variation within the group (SSW).

SSW is also known as the error variance. In notation:

$$SST = SSB + SSW$$

$$\sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \bar{\bar{Y}})^2 = \sum_{j=1}^{c} n_j(\bar{Y}_j - \bar{\bar{Y}})^2 + \sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2$$

Notations

SST: total sum of squares

SSB: the sum of squares between groups (between sum the of squares)

SSW: the sum of squares within groups

c: numbers of groups or levels

n_j: number of observations in group j

Y_{ij}: i^{th}observation from group j

$\bar{\bar{Y}}$: grand mean (mean of all data values) or overall mean

$\bar{Y}_j$: the sample mean from group j or group mean

Figure 3 provides a pictorial representation of observations belonging to three levels of a factor and their distribution with respect to the overall mean. Figure 4 depicts the differences between the group means and the overall mean.

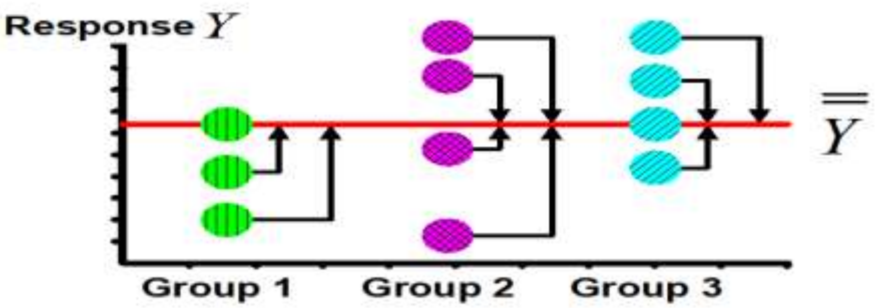

Figure 3: Representation of SST
(Courtesy: Malhotra, N., Hall, J., Shaw, M., & Oppenheim, P. (2006). Marketing research: An applied orientation. Pearson Education Australia.)

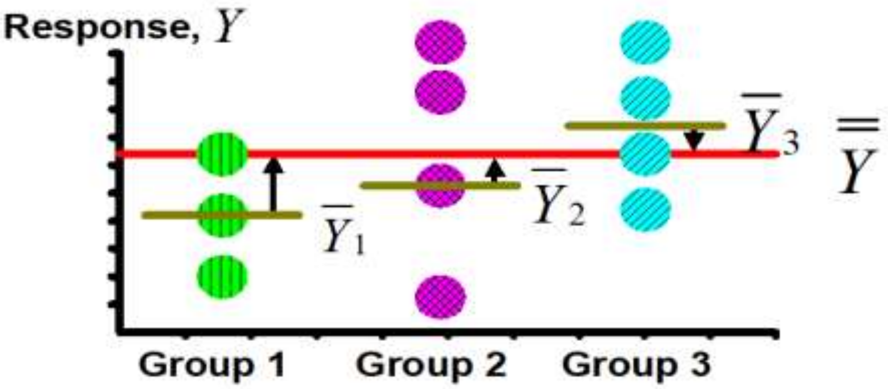

Figure 4: Representation of SSB
(Courtesy: Malhotra, N., Hall, J., Shaw, M., & Oppenheim, P. (2006). Marketing research: An applied orientation. Pearson Education Australia.)

If all observations are from the same population, then the group means would be equal, and all will be equal to the overall mean. In other words, under the null hypothesis, $\bar{Y}_j$, j = 1, …, c and $\bar{\bar{Y}}$ are close. They may not be identical because of the sampling fluctuations. As a result, we expect SSB = $\sum_{j=1}^{c} n_j(\bar{Y}_j - \bar{\bar{Y}})^2)$ to be small, since this quantifies the between group variance. A large value of SSB indicates that the population means are indeed different and the null hypothesis may not hold. Now let us consider the within group sum of square SSW = $\sum_{j=1}^{c}\sum_{i=1}^{n_j}(Y_{ij} - \bar{Y}_j)^2)$ which measure the variability within each group. A pictorial representation is given in Figure 5.

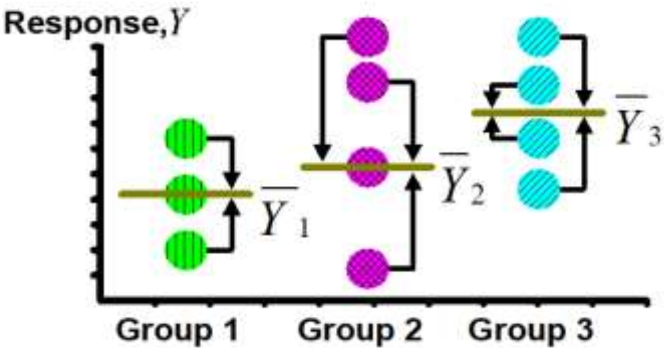

Figure 4: Representation of SSW

(Courtesy: Malhotra, N., Hall, J., Shaw, M., & Oppenheim, P. (2006). Marketing research: An applied orientation. Pearson Education Australia.)

If the null hypothesis holds in the population, then the total variability in the data correspond to SSW. If, however, the null hypothesis fails to hold, and the group means are all different, then most of the variability in the data is explained by SSB. Hence compared to SSB, SSW is expected to be small.

Note that, given a set of observations, SST is constant.

Another important concept is the Degrees of Freedom (DF). This equals the number of independent quantities contributing to construct the sums of squares. Each defined sum of squares has a corresponding DF. Dividing each SS by the appropriate DF; the mean sum of squares is obtained.

The ***mean sum of squares between groups*** is calculated as:

$$MSB = \frac{SSB}{c-1}$$

The ***mean of sum of the squares within groups*** is calculated as

$$MSW = \frac{SSW}{n-c}$$

A ratio between MSB and MSW provides an indication regarding whether the null hypothesis can be accepted or not. This ratio is defined as

$$F_{STAT} = \frac{MSB}{MSW}$$

According to the rationale explained above, if MSB is too large compared to MSW, the null hypothesis is rejected. F_{STAT} follows an F distribution with $df\ (c - 1, n - c)$. Since F_{STAT} is a ratio of two positive quantities, it is always positive. Hence the rejection rule is:

Reject $H_0\ if\ \ F_{STAT} > \ F_\alpha$. Figure 6 shows the right hand side critical region.

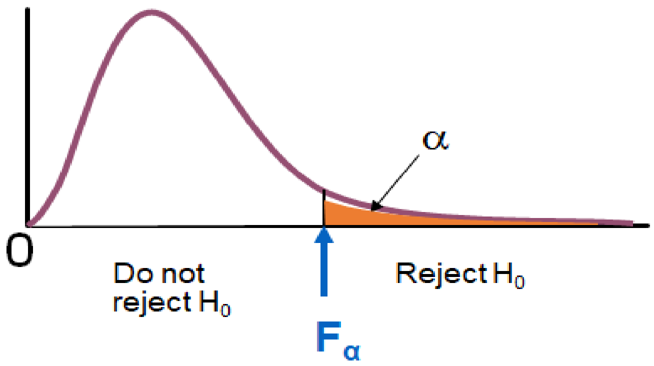

Figure 5: Critical region of F_{STAT}
(Courtesy: Malhotra, N., Hall, J., Shaw, M., & Oppenheim, P. (2006). Marketing research: An applied orientation. Pearson Education Australia.)

Results of ANOVA is presented in a tabular form as shown below.

Table 1: One-way ANOVA

Source of variation	Degrees of Freedom (df)	Sum of Squares	Mean Sum of Squares	F-test
Between groups	$c-1$	SSA	$MSA = \frac{SSA}{c-1}$	$F_{STAT} = \frac{MSA}{MSW}$
Within groups	$n-c$	SSW	$MSW = \frac{SSW}{n-c}$	
Total	$n-1$	SST		

Case Study continued.

Solution: The objective is to determine whether CO_2 emission from cars depends on fuel type or manufacturer or both.

Descriptive Analysis (EDA) on Car Data

```
#Step 1: Import important packages into Jupyter Notebook
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

#Step 2: Read the dataset into Jupyter Notebook using read_csv
aovData = pd.read_csv('AOVData.csv')

aovData.shape
(510, 4)

aovData.head()

      Car_ID manufacturer fuel_type co_emissions
0       1        Audi        Petrol      441.55
1       2        BMW         E85         376.47
2       3        BMW         E85         414.12
```

```
3         4         BMW          E85          351.41
4         5         Volvo        E85          284.59
```

Note that “co_emissions” is the response Y and “fuel_type” and “manufacturer” are two factors at multiple levels.

```
#Step 3: Summary of response: Carbon emission

aovData['co_emissions'].describe().transpose()

count  mean   std    min    25%    50%    75%     max
510.0 358.46 66.91 162.07 312.63 356.19 410.645 544.56
Name: co_emissions, dtype: float64

bin_edges = np.arange(160, 560, 20)
plt.hist(aovData.co_emissions,
         bins=bin_edges,
         density=False,
         histtype='bar',
         color='b',
         edgecolor='k',
         alpha=0.5);
```

The minimum value of carbon emission is 162.1 and maximum is 544.6 and the mean value is 358.5. Using figure (7) and figure (8), the pattern of carbon emission can be visualized.

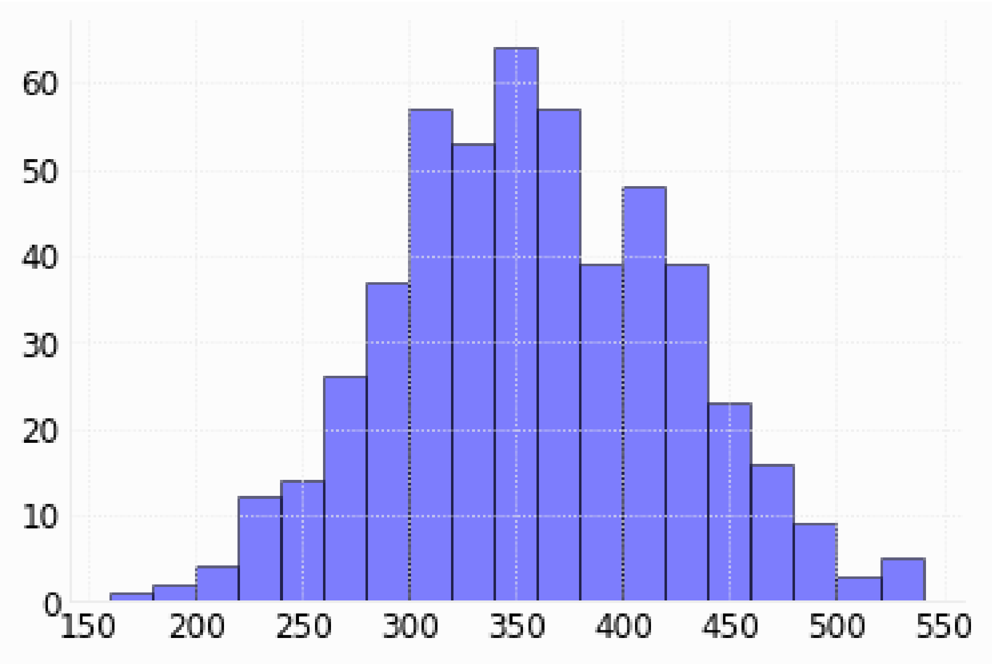

Figure 6: Histogram of Carbon Emission

```
sns.boxplot(aovData['co_emissions'] , orient = 'v')
plt.show()
```

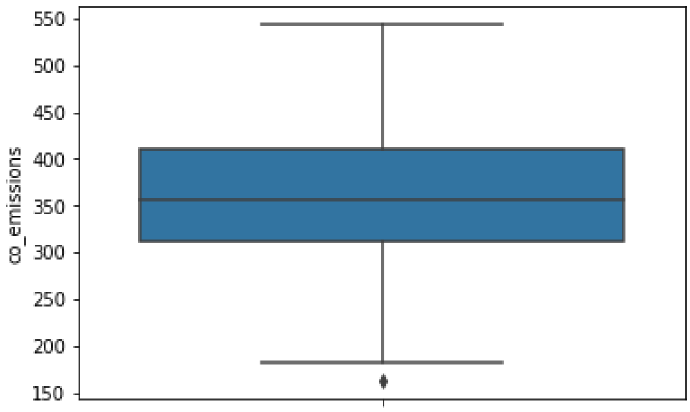

Figure 7: Box plot of Carbon Emission

Frequency counts and mean of carbon emission at different levels of the factors are shown below.

```
# #Factor 1: fuel_type

aovData['fuel_type'].value_counts()

Petrol LPG E85
179    170 161
Name: fuel_type, dtype: int64

aovData.groupby("fuel_type")["co_emissions"].mean()

E85     LPG    Petrol
338.12 363.74 371.72
Name: co_emissions, dtype: float64
```

```
#Factor 2: manufacturer

aovData['manufacturer'].value_counts()

Audi Ford Volvo BMW
142  132  123   113
Name:manufacturer, dtype: int64

aovData.groupby("manufacturer")["co_emissions"].mean()

Audi    Ford    Volvo   BMW
349.73  377.54  365.08 343.90
Name: co_emissions, dtype: float64
```

<u>Problem 1: Whether there is any dependency on Y of X_1: Fuel Type</u>

We need to test the hypothesis that the use of three different fuel types does not impact carbon emission. Formally

H_0: $\mu_1 = \mu_2 = \mu_3$ against H_a: At least one carbon emission level is different from the rest.

Before one-way ANOVA procedure is applied to the data, visual comparison is recommended. Moreover, the normality and equality of variance assumptions need to be checked.

```
a4_dims = (7,7)
fig, ax = plt.subplots(figsize=a4_dims)
a = sns.boxplot(x= "fuel_type", y = 'co_emissions' , data = aovData, hue
= 'fuel_type')
a.set_title("Carbon Emission w.r.t. Fuel type (3 levels)",fontsize=15)
plt.show()
```

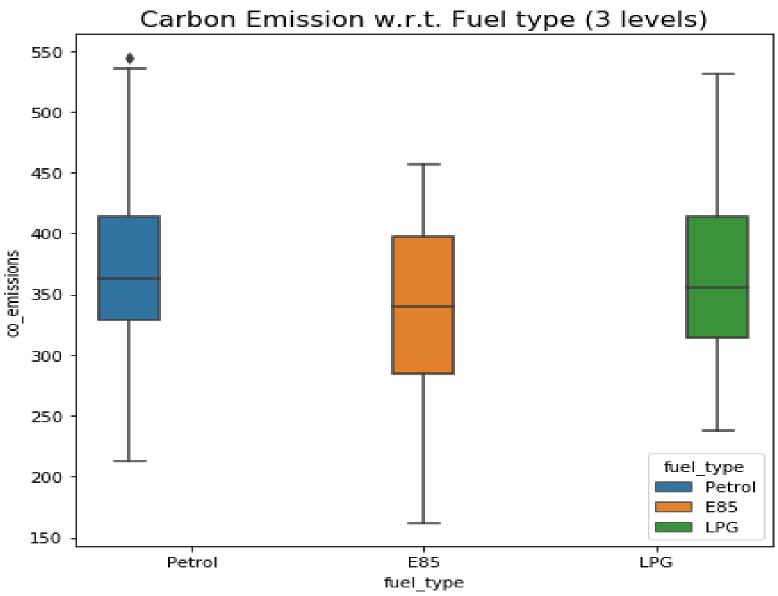

Figure 8: Carbon Emission w.r.t. Fuel type (3 levels)

For testing of normality, Shapiro-Wilk's test is applied to the response.

H_0: Carbon emission follows a normal distribution against H_a: Carbon emission does not follow a normal distribution

```
#Aussmption 1: Normality

from scipy import stats

w, p_value = stats.shapiro(aovData['co_emissions'])

print("W = {}".format(w), "p_value = {}".format(p_value))
W = 0.997 p_value = 0.4972
```

Since p-value of the test is very large, we fail to reject the null hypothesis that the response follows the normal distribution.

Next, we need to test the assumption that at all three levels of the factor fuel_type, population variance is equal. In other words, the homogeneity of variance assumption is satisfied. We may formulate the problem as:

$H_0: \sigma_1 = \sigma_2 = \sigma_3$ against H_a: At least one variance is different from the rest.

```
#Aussmption 2: Homogeneity of Variance

statistic, p_value = stats.levene(

             aovData['co_emissions'][aovData['fuel_type']=="Petrol"],

             aovData['co_emissions'][aovData['fuel_type']=="E85"],

             aovData['co_emissions'][aovData['fuel_type']=="LPG"])


print("statistic = {}".format(statistic), "p_value = {}".format(p_value)
)
statistic = 1.643 p_value = 0.194
```

Since the p-value is large, we fail to reject the null hypothesis of homogeneity of variances.

Once the two assumptions of one-way ANOVA are satisfied, we can now compare the population means.

```
#Apply one-way ANOVA

mod = ols('co_emissions ~ fuel_type', data = aovData).fit()
aov_tbl = sm.stats.anova_lm(mod, type = 1)
print(aov_tbl)

              df        sum_sq       mean_sq          F    PR(>F)
fuel_type    2.0  1.028130e+05  51406.481215  11.976652  0.000008
Residual   507.0  2.176158e+06   4292.224647        NaN       NaN
```

Let us consider the summary output known as ANOVA Table.

For the given problem sum of squares due to the factor fuel_type (SSB) is 102813 and the sum of squares due to error (SSW) is 2176158. The total sum of squares (SST) for the data is (102813+2176158=2278971). Since the factor has 3 levels, DF corresponding to fuel_type is 3 – 1 = 2. Total DF is 510 – 1 = 509. Hence DF due to error is 509 – 2 = 507. Mean sum of squares is obtained by dividing the sums of squares by corresponding DF. The value of the F-statistic is approximately 12 and the p-value is highly significant.

Based on the ANOVA test we, therefore, reject the null hypothesis that the three population means are identical. At least for one fuel-type mean carbon emission is different from the rest.

Residuals are defined as the difference between the observed values and the expected values. Detail discussion on residuals will be taken up along with Simple Linear Regression. The following two graphs are introduced to check the distribution of the residuals. Fig 10 indicates that the mean carbon emission of two groups are close but the other group means stands out. It also supports the homoscedasticity of variances. Fig 11 indicates that the normality assumption holds.

```
# model values
model_fitted_y = mod.fittedvalues
# model residuals
model_residuals = mod.resid
# normalized residuals
model_norm_residuals = mod.get_influence().resid_studentized_internal
# absolute squared normalized residuals
model_norm_residuals_abs_sqrt = np.sqrt(np.abs(model_norm_residuals))
# absolute residuals
model_abs_resid = np.abs(model_residuals)
# leverage, from statsmodels internals
model_leverage = mod.get_influence().hat_matrix_diag
# cook's distance, from statsmodels internals
model_cooks = mod.get_influence().cooks_distance[0]

plot_lm_1 = plt.figure()
plot_lm_1.axes[0] = sns.residplot(model_fitted_y, 'co_emissions', data=a
ovData,
                          lowess=True,
                          scatter_kws={'alpha': 0.5},
                          line_kws={'color': 'red', 'lw': 1, 'alpha': 0.
8})

plot_lm_1.axes[0].set_title('Residuals vs Fitted')
plot_lm_1.axes[0].set_xlabel('Fitted values')
plot_lm_1.axes[0].set_ylabel('Residuals')
```

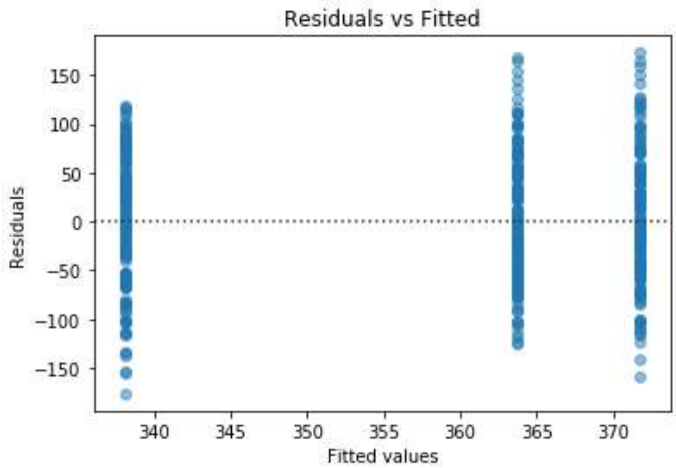

Figure 9: Residuals vs. Fitted plot w.r.t. Fuel Type

```
QQ = ProbPlot(model_norm_residuals)
```

```
plot_lm_2 = QQ.qqplot(line='45', alpha=0.5, color='#4C72B0', lw=1)
plot_lm_2.axes[0].set_title('Normal Q-Q')
plot_lm_2.axes[0].set_xlabel('Theoretical Quantiles')
plot_lm_2.axes[0].set_ylabel('Standardized Residuals');

# annotations
abs_norm_resid = np.flip(np.argsort(np.abs(model_norm_residuals)), 0)
abs_norm_resid_top_3 = abs_norm_resid[:3]
for r, i in enumerate(abs_norm_resid_top_3):
    plot_lm_2.axes[0].annotate(i,xy=(np.flip(QQ.theoretical_quantiles, 0
)[r],model_norm_residuals[i]))
```

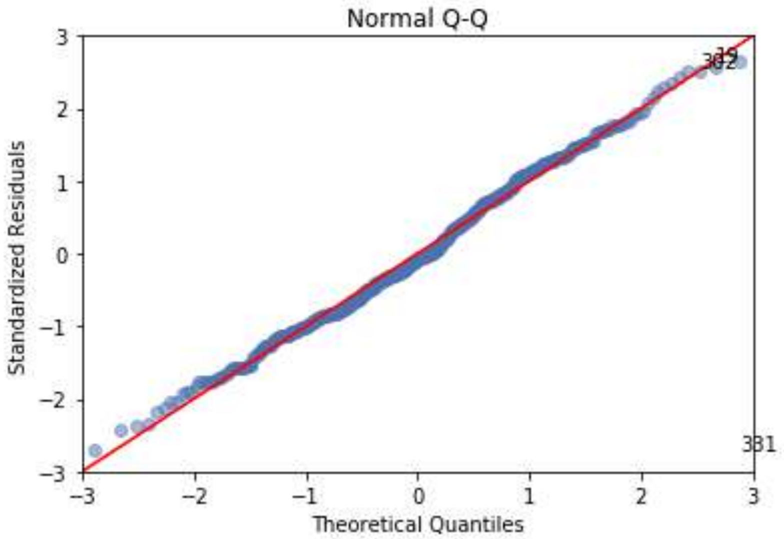

Figure 10: Normal Q-Q plot w.r.t. Fuel Type

Note that once the null hypothesis of equality of means is rejected, the next natural question is to find out which mean(s) is different from the rest. Before we answer that question, let us first check whether carbon emission is dependent on the manufacturer.

Problem 2: Whether there is any dependency on Y of X_2: Manufacturer

We need to test the hypothesis that carbon emission is the same for all car manufacturer. Formally,

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ against H_a: At least for one manufacturer emission level is different from the rest.

As in the previous problem, visual comparison of group means is recommended.

```
a4_dims = (7,7)
fig, ax = plt.subplots(figsize=a4_dims)
a = sns.boxplot(x= "manufacturer", y = 'co_emissions' , data = aovData,
hue = 'manufacturer')
a.set_title("Carbon Emission w.r.t. manufacturer (4 manufacturer)",fonts
ize=15)
```

```
plt.show()
```

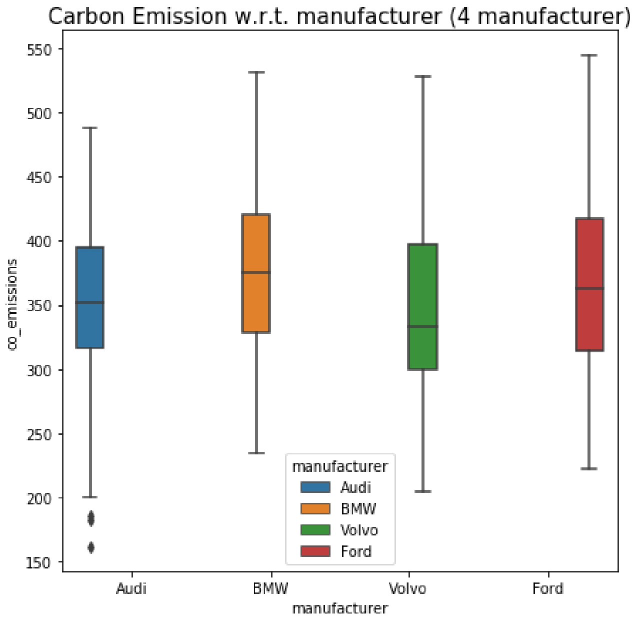

Figure 11: Carbon Emission w.r.t. manufacturer (4 levels)

Assumption 1 has already been tested for this data.

Equality of variance assumption needs to be checked for this factor.

In order to test the assumption that for all four manufacturers, population variance is equal to the following null and alternative hypothesis are defined as:

$H_0\!: \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$ against H_a: At least one variance is different from the rest.

```
statistic, p_value = stats.levene(
                aovData['co_emissions'][aovData['manufacturer']=="Audi"],
                aovData['co_emissions'][aovData['manufacturer']=="BMW"],
               aovData['co_emissions'][aovData['manufacturer']=="Volvo"],
               aovData['co_emissions'][aovData['manufacturer']=="Ford"])
print("statistic = {}".format(w), "p_value = {}".format(p_value))

statistic = 0.997 p_value = 0.201
```

Since the p-value is large, we fail to reject the null hypothesis of homogeneity of variances and can say that population variances are equal across different manufacturers. As all the assumptions of one-way ANOVA are satisfied, we can now compare the population means with respect to the manufacturer.

ANOVA Table for manufacturer

```
#Apply one-way ANOVA

mod = ols('co_emissions ~ manufacturer', data = aovData).fit()

aov_tbl = sm.stats.anova_lm(mod, type = 1)

print(aov_tbl)
                    df        sum_sq       mean_sq         F    PR(>F)

manufacturer     3.0  8.382457e+04  27941.524040  6.44076  0.000276

Residual       506.0  2.195146e+06   4338.233767       NaN       NaN
```

For the given problem sum of squares due to the manufacturer (SSB) is 83825 and the sum of squares due to error (SSW) is 2195146. The total sum of squares (SST) for the data is (83825+2195146=2278971). Since the factor has 4 levels, DF corresponding to the manufacturer is 4 – 1 = 3. Total DF is 510 – 1 = 509. Hence DF due to error is 509 – 3 = 506. Mean sum of squares is obtained by dividing the sums of squares by corresponding DF. The value of the F-statistic is approximately 6 and the p-value is highly significant.

Therefore, based on the ANOVA test, we reject the null hypothesis that the four population means are the same. At least for one manufacturer mean carbon emission is different from the rest.

Two important points need to be noted here

1) Whether we are testing equality of mean across fuel type or manufacturer, SST is constant given data. In this case SST = 2278971.
2) Total DF is constant given a data and is equal to n – 1. Since sample size is 510, total DF = 509.

Residual plots are shown below for different manufacturers.

```
# model values
model_fitted_y = mod.fittedvalues
# model residuals
model_residuals = mod.resid
# normalized residuals
model_norm_residuals = mod.get_influence().resid_studentized_internal
# absolute squared normalized residuals
model_norm_residuals_abs_sqrt = np.sqrt(np.abs(model_norm_residuals))
# absolute residuals
model_abs_resid = np.abs(model_residuals)
# leverage, from statsmodels internals
model_leverage = mod.get_influence().hat_matrix_diag
# cook's distance, from statsmodels internals
model_cooks = mod.get_influence().cooks_distance[0]

plot_lm_1 = plt.figure()
plot_lm_1.axes[0] = sns.residplot(model_fitted_y, 'co_emissions', data=a
ovData,
                       lowess=True,
```

```
                                scatter_kws={'alpha': 0.5},
                                line_kws={'color': 'red', 'lw': 1, 'alpha': 0.8})

plot_lm_1.axes[0].set_title('Residuals vs Fitted')
plot_lm_1.axes[0].set_xlabel('Fitted values')
plot_lm_1.axes[0].set_ylabel('Residuals')
```

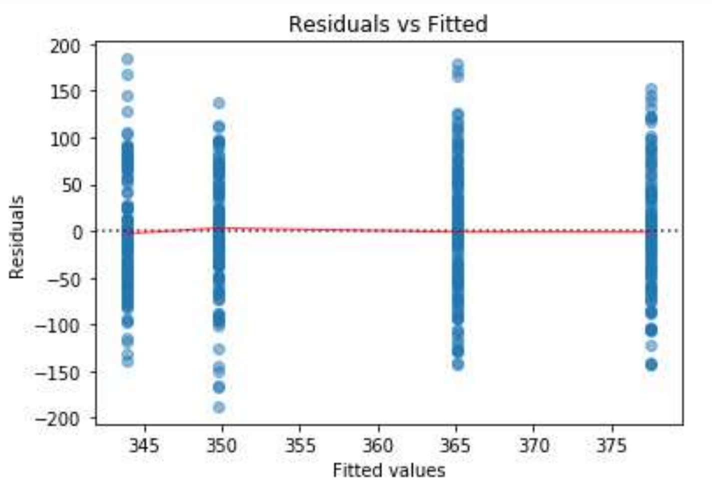

Figure 12: Residuals vs. Fitted plot w.r.t. manufacturer

```
QQ = ProbPlot(model_norm_residuals)
plot_lm_2 = QQ.qqplot(line='45', alpha=0.5, color='#4C72B0', lw=1)
plot_lm_2.axes[0].set_title('Normal Q-Q')
plot_lm_2.axes[0].set_xlabel('Theoretical Quantiles')
plot_lm_2.axes[0].set_ylabel('Standardized Residuals');
# annotations
abs_norm_resid = np.flip(np.argsort(np.abs(model_norm_residuals)), 0)
abs_norm_resid_top_3 = abs_norm_resid[:3]
for r, i in enumerate(abs_norm_resid_top_3):
    plot_lm_2.axes[0].annotate(i,xy=(np.flip(QQ.theoretical_quantiles, 0
)[r],model_norm_residuals[i]))
```

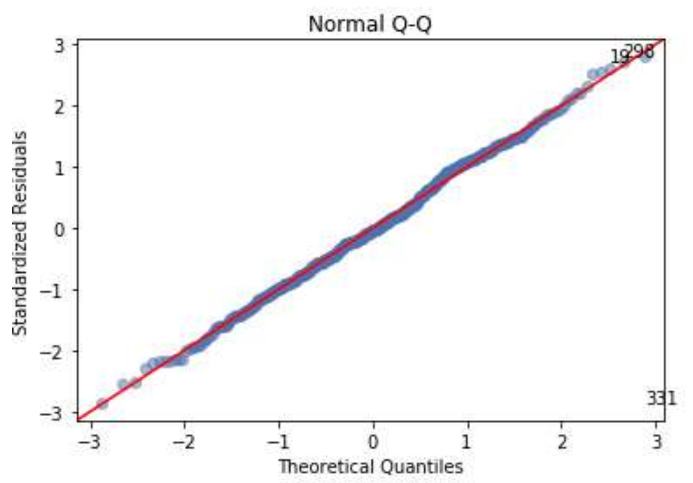

Figure 13: Normal Q-Q plot w.r.t. Fuel Type

3.3 Multiple Comparison Test: The Tukey's HSD and Tukey-Kramer procedure

We have observed that Fuel_type and Manufacturer individually have a significant impact on Carbon emission as null hypotheses that group means are equal have been rejected in both cases. However, we have not been able to determine which mean is different from the rest or whether all pairs of means are different. There are special tests (called ***post hoc tests***) of the differences between all pairs of means. These tests are also called multiple comparison tests.

These tests are NOT independent t-tests, because here ALL pairs of group means are considered simultaneously.

Before we can introduce the multiple comparison tests, we need to discuss about an adjustment to the Type I error of this test.

Type I error, α is the probability of rejecting a null hypothesis when it is true. Hence the probability of accepting a null hypothesis when it is true is:

$1 - \alpha$ = 1-(.05) = 0.95

Consider now two independent null hypotheses, both are being tested at level α. Both null hypotheses will be accepted, when both are indeed true is (0.95)*(0.95) = 0.9075, by application of probability multiplication rule. Hence, the probability of making Type I error in this case is 1 – 0.9075 = 0.0975. Note that, even though at individual test level, Type I error had been fixed at α = 0.05, in effect, because of two null hypotheses being tested simultaneously, level of the test has increased, i.e. probability of rejecting at least one null hypothesis, when it is actually true is higher than the fixed value. As more and more null hypotheses will be tested simultaneously, Type I error rate will keep on inflating with increasing number of tests as shown in Figure 15. The ***family-wise error rate*** is the probability that at least one type I error is made on a set of tests.

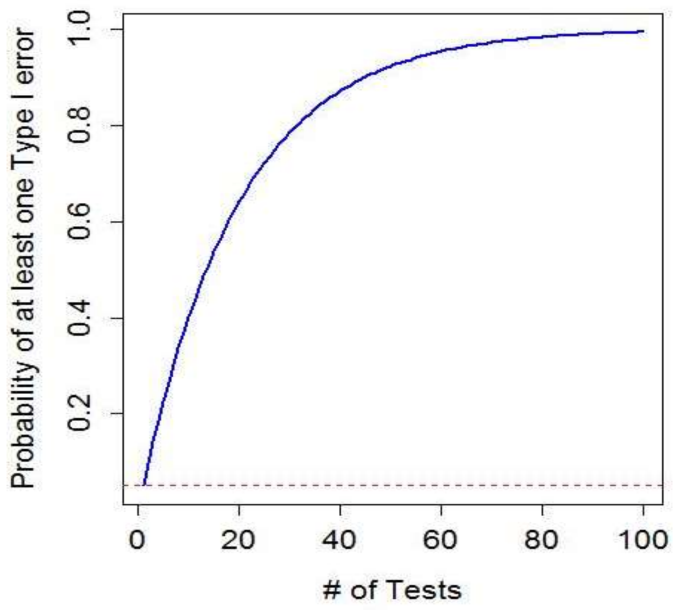

Figure 14: Probability of inflating type 1 error
(Courtesy: Greenwood, M. and Banner, K., "Multiple (pair-wise) comparisons using Tukey's HSD and the compact letter display")

The most important part of multiple comparison test is an adjustment for α.

Essentially multiple comparison methods consider each pair of group mean to see whether their difference is significant or not. If a factor has 3 levels (e.g. fuel type), then the number of pairs compared is $\binom{3}{2}$ = 3; if a factor has 4 levels, (e.g. manufacturer), then the number of pairs compared is $\binom{4}{2}$ = 6. As levels of a factor increases, the number of pairs increases accordingly and we should we worried about inflated family-wise error rate. If ANOVA F-test is rejected, then at least one of the pairs will be found significant.

There are several tests available for multiple comparisons which are introduced below.

Tukey's HSD is recommended when sample sizes are equal, or approximately equal for each level of the factor, i.e. $n_j = \frac{n}{c}$, where c is the number of factor levels. A modification of Tukey's HSD is suggested by C.Y. Kramer to accommodate unequal group sizes. Another often used procedure is Bonferroni procedure.

Case Study continued.

Multiple comparison tests for X_1: Fuel Type

In order to identify for which fuel type mean carbon emission is different from other groups, the hypotheses may be stated as:

H_0: All pairs of group means are equal against H_a: At least one group mean is different from the rest.

In this case, as there are only 3 pairs to be considered, we may write the null and alternative hypothesis as:

$H_0: \mu_1 = \mu_2$ and $\mu_1 = \mu_3$ and $\mu_2 = \mu_3$ against H_a: $\mu_1 \neq \mu_2$ or $\mu_1 \neq \mu_3$ or $\mu_2 \neq \mu_3$ respectively, where $\mu_1$represents mean carbon emission when fuel type is E85, μ_2 represents mean carbon emission when fuel type is LPG and μ_3 is the same for Petrol.

```
# Posthoc test: Tukey test

MultiComp=MultiComparison(aovData['co_emissions'],aovData['fuel_type'])

print(MultiComp.tukeyhsd().summary())

 Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower   upper  reject
----------------------------------------------------
   E85    LPG  25.6199 0.0012  8.6837  42.556   True
   E85 Petrol  33.5984  0.001 16.8707 50.3262   True
   LPG Petrol   7.9785 0.4931 -8.5144 24.4715  False
----------------------------------------------------
```

P-value is significant for comparing carbon emission mean levels for the pair LPG-E85 and Petrol-E85, but not for Petrol-LPG. The null hypothesis of equality of all population means is rejected. It is now clear that mean carbon emission for Petrol and LPG is similar but emission for fuel type E85 is significantly different from these two.

Note also that, the numerical values of the differences being positive, mean carbon emission for fuel type E85 is significantly lower than that for petrol or LPG. This same observation is borne out by the residual plot in Fig 10, where the values of the residuals corresponding to E85 is lower compared to the other fuel types, which are much closer.

Often it is easier to visualize the difference among group means.

```
#for family wise comparison

results = MultiComp.tukeyhsd()
df=results.summary()
results_as_html = df.as_html()
df1=pd.read_html(results_as_html, header=0, index_col=0)[0].reset_index(
)
groups = np.array([df1.group1+ '-'+ df1.group2])

plt.figure(figsize=(8,7))
data_dict = {}
data_dict['category'] = groups.ravel()
data_dict['lower'] = results.confint[:,0]
data_dict['upper'] = results.confint[:,1]
dataset = pd.DataFrame(data_dict)

for lower,upper,y in zip(dataset['lower'],dataset['upper'],range(len(dat
aset))):
    plt.plot((lower,upper),(y,y),'ro-',color='orange')
plt.yticks(range(len(dataset)),list(dataset['category']));
```

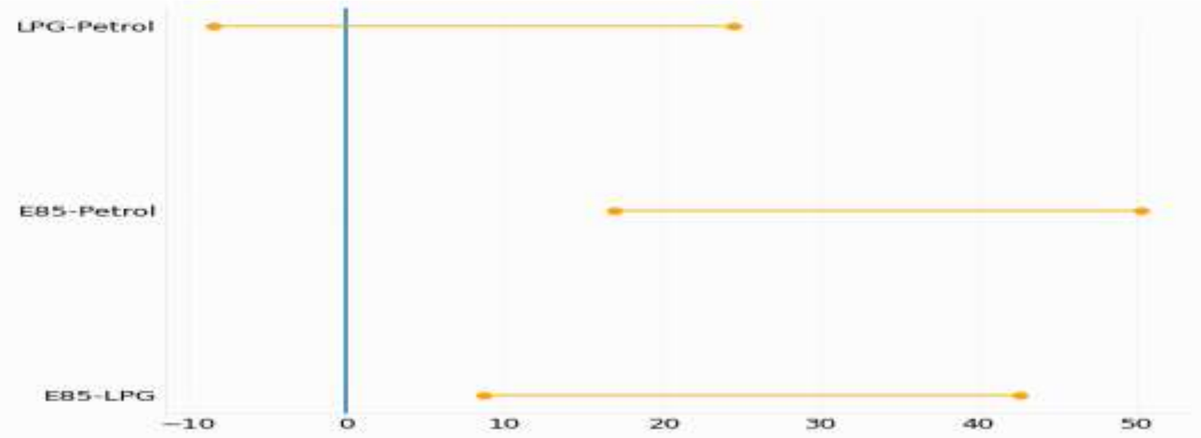

Figure 15: Family-wise comparison for fuel type

Figure 16 is a graphical representation of pair-wise comparisons from Tukey's HSD for fuel type. The confidence intervals not containing 0 is for the difference between LPG & E85 and for the difference between Petrol & E85. This indicates that population means of these pairs of fuels are different. From the values of the pairwise differences, it may also be concluded that carbon emission from cars using E85 is significantly less than the other two.

Let us now determine cars by which manufacturer have a mean carbon emission level different from the others.

Multiple comparison tests for X_2: Manufacturer

In order to identify for which manufacturer mean carbon emission is different from others, the hypotheses may be stated as:

H_0: All pairs of group means are equal against H_a: At least one group mean is different from the rest.

We may also rewrite the null and alternative hypotheses as

$H_0: \mu_i = \mu_j$ against $H_a: \mu_i \neq \mu_j$, for all $i \neq j$, $i, j = 1, 2, 3, 4$. Subscript 1 represents Audi, resents mean value of 2 BMW, 3 Ford and 4 Volvo.

```
## post hoc test

MultiComp=MultiComparison(aovData['co_emissions'],aovData['manufacturer'])

print(MultiComp.tukeyhsd().summary())

Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2 meandiff p-adj   lower    upper   reject
-----------------------------------------------------
  Audi    BMW  27.8115 0.0048   6.4089  49.2141   True
  Audi   Ford  15.3513 0.2178  -5.1756  35.8782  False
  Audi  Volvo   -5.829  0.882 -26.7415  15.0835  False
   BMW   Ford -12.4602 0.4541 -34.2191   9.2987  False
```

```
   BMW  Volvo -33.6405  0.001 -55.7635 -11.5175   True
  Ford  Volvo -21.1803 0.0516 -42.4573   0.0967  False
------------------------------------------------------
```

It is clear from the above table that there is a significant difference in mean carbon emission between (i) BMW and Audi; between (ii) Volvo and BMW and between (iii) Volvo and Ford.

```
df=results.summary()
results_as_html = df.as_html()
df1=pd.read_html(results_as_html, header=0, index_col=0)[0].reset_index(
)
groups = np.array([df1.group1+ '-'+ df1.group2])



plt.figure(figsize=(10,10))
data_dict = {}
data_dict['category'] = groups.ravel()
data_dict['lower'] = results.confint[:,0]
data_dict['upper'] = results.confint[:,1]
dataset = pd.DataFrame(data_dict)

for lower,upper,y in zip(dataset['lower'],dataset['upper'],range(len(dat
aset))):
    plt.plot((lower,upper),(y,y),'ro-',color='orange')
plt.yticks(range(len(dataset)),list(dataset['category']));
```

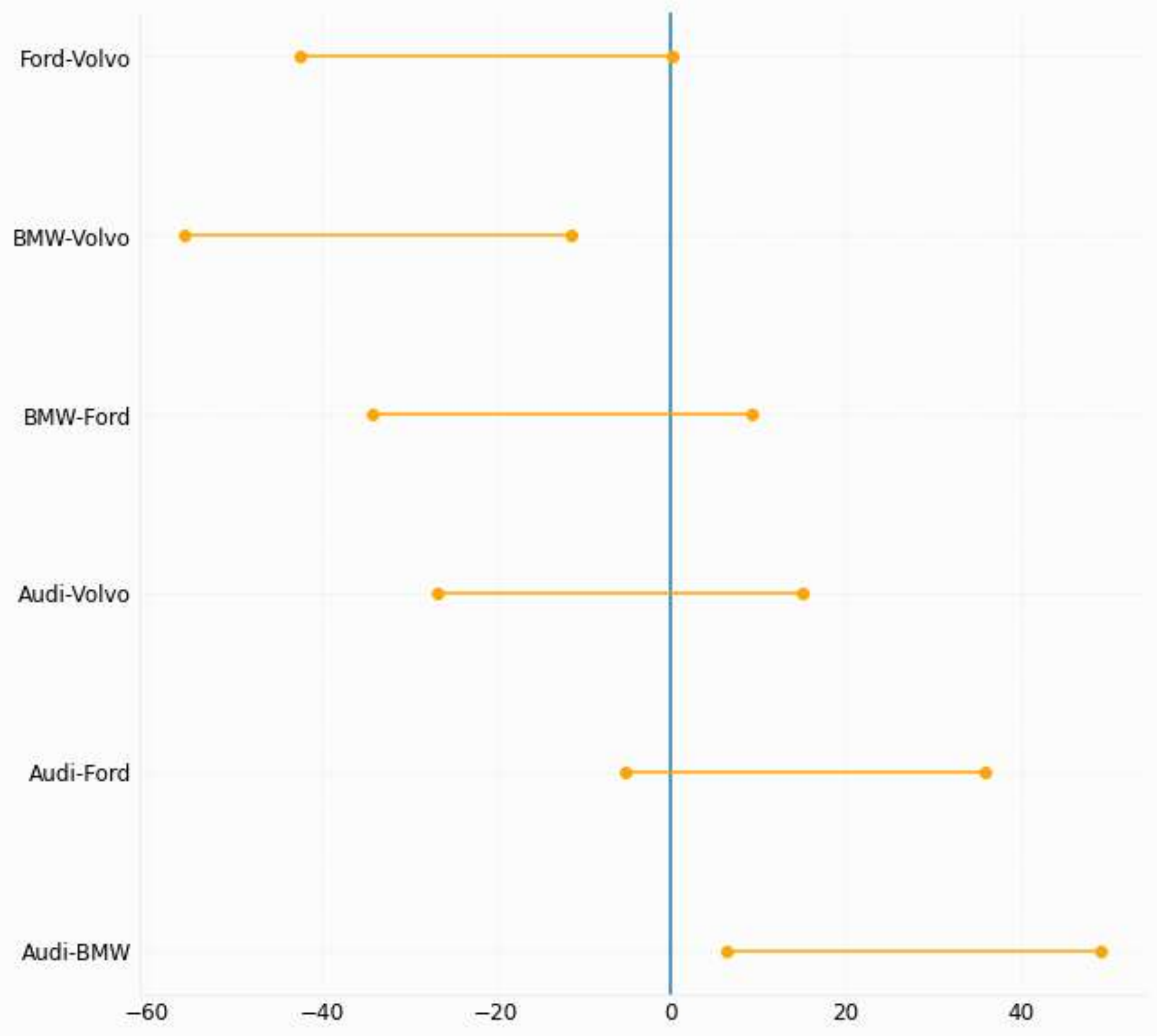

Figure 16: Family-wise comparison for Manufacturer

Here, using Figure 17, it is evident that the confidence intervals of difference of mean carbon emission for the pairs BMW and Audi, Volvo and BMW, and Volvo and Ford do not contain zero, therefore these pairs are significantly different from each other.

Important points to note regarding assumptions on ANOVA

ANOVA procedure is sensitive on both normality and homogeneity of variance assumptions. For test of normality, we have started with overall normality test. In this case, the null hypothesis of normality is not rejected. However, there will be situations when overall normality test will be rejected. If so, then normality test must be carried out within each group. If a factor has K levels, then K separate tests of normality must be performed and all K distributions must be normal.

In case where at least in one group normality assumption is violated, a non-parameteric alternative to ANOVA must be employed. This is known as Kruskal-Walli's test. This test depends on the ranking of the observations, rather than the values of the observations.

Homogeneity of variance assumption is also a very important assumption. If the group variances are not assumed equal, then the F-test for ANOVA is not defined. Recall also that the MSE = SSE/df is an estimate of the variance. If the equality of variance is rejected and if the groups have equal sample sizes, then F-test is a robust test; otherwise the test is biased with inflated Type I error probability.

4. Two-way ANOVA

In practical scenario almost never only one factor is studied in isolation for its effect on the response. In case of one-way ANOVA, one does not have the flexibility to evaluate how responses to one treatment behave with respect to the levels of other treatments.

Therefore, multi-factor experiments are extremely common and preferable in practice.

4.1 Two-way Analysis of Variance

Two-way ANOVA uses two factors (independent or interacting) to test various hypotheses of interest. With two factors we may think of a contingency table with the levels of the two factors making up the rows and the columns. The system of notation is rather complex. Let the two factors be denoted by A and B, levels of one will be in the row and levels of the other will be in the column of the contingency table.

Before we specify the hypothesis, the notion of interaction needs to be introduced.

Interaction is a quantification of association of two factors. If one factor behaves differently at different levels of one or more factors, an interaction effect is said to exist.

Interaction term may have different names in various fields. For example,

In medicine, the doctor always asks what other medications a patient is on before prescribing a new medication so that either the two medicines do not impact each other or both together may become more efficacious.

Interaction occurs when the pattern of the cell means in one row (going across columns) varies from the patterns of cell means in other rows. Graphically it can be shown in Figure 18.

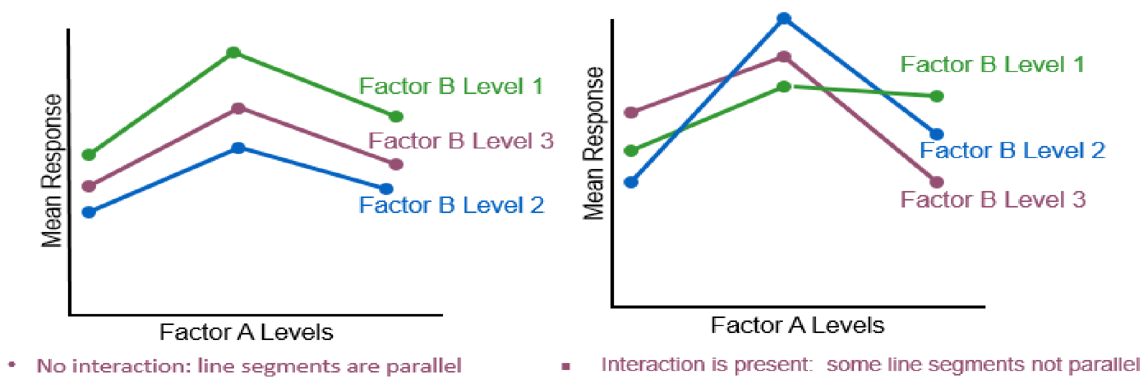

Figure 17: Interaction among factors
(Courtesy: Malhotra, N., Hall, J., Shaw, M., & Oppenheim, P. (2006). Marketing research: An applied orientation. Pearson Education Australia.)

Notations:

r: number of levels of Factor A, (row), levels denoted by $i = 1 \ldots r$

c: number of levels of Factor B, (column), levels denoted by $j = 1 \ldots c$

Each cell in the contingency table may have more than one observation. Let n_{ij} denote the number of observations (replications) in $(i, j)^{th}$ cell. If Y_{ijk} denotes the k^{th}observation in the $(i, j)^{th}$ cell, then the mean of all the observations in that cell will be denoted by $\bar{Y}_{ij.}$: $\frac{\sum_{k=1}^{n_{ij}} Y_{ijk}}{n_{ij}}$.

Similarly we can define the row means and the column means, denoted by $\bar{Y}_{i..}$ and $\bar{Y}_{.j.}$ respectively, for all i and j. Formally

$\bar{Y}_{i..}$: $\frac{\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}} Y_{ijk}}{\sum_{j=1}^{c} n_{ij}}$ and $\bar{Y}_{.j.}$: $\frac{\sum_{i=1}^{r}\sum_{k=1}^{n_{ij}} Y_{ijk}}{\sum_{i=1}^{r} n_{ij}}$

The overall mean or the Grand Mean is denoted by $\bar{\bar{Y}}$, which is the sample mean of all the observations.

Case Study continued.

We are now interested in studying the impact of both fuel type and manufacturer on carbon emission of the cars. If fuel type is assumed to be the row factor and manufacturer, the column factor, *r* = 3 and *c* = 4.

The table below shows the replications as well as the cell means for each cell of the contingency table. The row means, the column means and the grand mean are also provided.

Table 2: Carbon Emission at each combination of Fuel Type and Manufacturer

Manufacturer	Audi		BMW		Ford		Volvo			
Fuel Type	Count	Cell Mean	Count	Cell Mean	Count	Cell Mean	Count	Cell Mean	Row Total Count	Row Mean
E85	47	330.3	34	339.7	41	337.3	39	347.0	161	338.1
LPG	47	356.1	35	382.2	46	379.8	42	339.3	170	363.7
Petrol	48	362.5	44	403.1	45	375.3	42	345.6	179	371.7
Column Total Count/Column Mean	142	349.7	113	377.5	132	365	123	343.9	**510**	**358.5**

Let us now formally introduce the hypotheses for two-way ANOVA.

Table 3: Hypothesis for Two-way ANOVA

Factor A (row) effects:	$H_0: \mu_{1..} = \mu_{2..} = \mu_{3..} = \cdots = \mu_{r..}$ $H_1: Not\ all\ \mu_{i..}$ are equal
Factor B (column) effects:	$H_0: \mu_{.1.} = \mu_{.2.} = \mu_{.3.} = \cdots = \mu_{.c.}$ $H_1: Not\ all\ \mu_{.j.}$ are equal
Interaction effects:	$H_0: The\ interaction\ effect\ does\ not\ exist$ $H_1: An\ interaction\ effect\ exists$

Two-way ANOVA follows the same assumptions as one-way ANOVA as discussed in section 3.1.

Assumptions

1. Populations are normally distributed. (Use Shapiro-Wilk test)
2. Populations have equal variances. (Use Levene's test)
3. Samples are randomly and independently drawn.

4.2 Partition of Variation in Two-way ANOVA?

In two-way ANOVA, total variation is divided into four additive and independent components.

- Sum of squares due to Factor A: SSA
- Sum of squares due to Factor B: SSB
- Sum of squares due to interaction: SSAB
- the sum of squares errors (SSE).

$$SST = SSA + SSB + SSAB + SSE$$

Total sum of squares ($SST = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \bar{\bar{Y}})^2$): Sum of squares of variation is calculated from grand mean and the individual data values.

Sum of squares of Factor A ($SSA = c\sum_{i=1}^{r} n_{i.}(\bar{Y}_{i..} - \bar{\bar{Y}})^2$): Row variation is calculated using grand mean and row means calculated at each row level of Factor A, where $n_{i.} = \sum_{j=1}^{c} n_{ij}$.

Sum of squares of Factor B ($SSB = r\sum_{j=1}^{c} n_{.j}(\bar{Y}_{.j.} - \bar{\bar{Y}})^2$): Column variation is calculated using grand mean and column means calculated at each column level of Factor B, where $n_{.j} = \sum_{i=1}^{r} n_{ij}$.

Sum of squares of Interaction ($SSAB = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}(\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{\bar{Y}})^2$): *Interaction* occurs when the effects of treatment vary according to the levels of treatment of the other effect or in other words it can be defined as the failure of the response to one factor to be the same at different levels of another factor.

Sum of squares of Errors ($SST = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \bar{Y}_{ij.})^2$): It is the mean difference of each individual values with respect to their cell mean value.

Results of ANOVA is provided in a table as below.

Table 4: Two-way ANOVA

Source of variation	Degrees of Freedom (df)	Sum of Squares	Mean Squares (Variance)	F-test
Factor A	$r-1$	SSA	$MSA = \frac{SSA}{r-1}$	$F_{STAT} = \frac{MSA}{MSE}$
Factor B	$c-1$	SSB	$MSB = \frac{SSB}{c-1}$	$F_{STAT} = \frac{MSB}{MSE}$
AB (interaction)	$(r-1)(c-1)$	$SSAB$	$MSAB = \frac{SSAB}{(r-1)(c-1)}$	$F_{STAT} = \frac{MSAB}{MSE}$
Error	$\sum_{i=1}^{r}\sum_{j=1}^{c}(n_{ij}-1)$	SSE	$= \frac{SSE}{\left(\sum_{i=1}^{r}\sum_{j=1}^{c}(n_{ij}-1)\right)}$	
Total	$n-1$	SST		

Case Study continued.

Problem 3: whether there is any dependency on Y of X_1 and X_2 (Fuel type and Manufacturer) together

Before two-way ANOVA procedure is applied to the data, descriptive analysis is done.

```
pd.crosstab(aovData['fuel_type'],aovData['manufacturer'])

##            Audi  BMW  Ford  Volvo
##    E85       47   34    41     39
##    LPG       47   35    46     42
##    Petrol    48   44    45     42

aovData.groupby(['fuel_type','manufacturer'])['co_emissions'].mean()

##            Audi    BMW    Ford   Volvo
## E85     330.33 339.70 337.32 346.99
## LPG     356.11 382.22 379.81 339.30
## Petrol  362.47 403.06 375.31 345.62
```

```
fig, axes = plt.subplots()
fig.set_size_inches(10,10)
a = sns.boxplot(data = aovData, y = "co_emissions", x = "fuel_type" , hu
e = 'manufacturer', orient = "v")
```

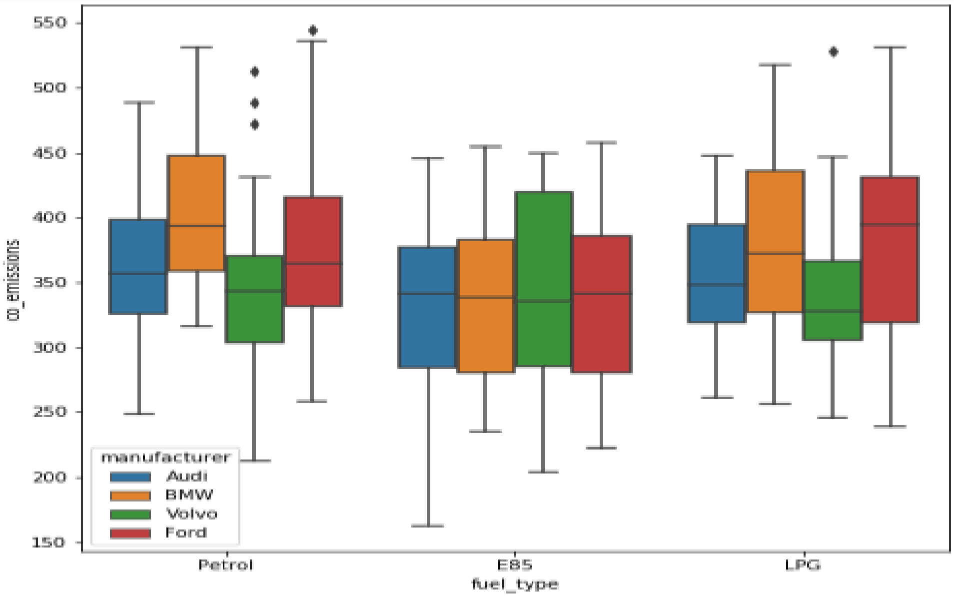

Figure 18: Box plot using fuel_type and manufacturer factors

To visualize interaction effect, a graphical representation of mean fuel_type across different manufacturers is shown in Figure 20. It is observed that the lines are not parallel. Mean carbon emission is lowest with E85 for all cars except BMW. This indicates interaction effect between fuel_type and manufacturer.

```
fig, ax = plt.subplots(figsize=(10, 6))
fig = interaction_plot(x=aovData['fuel_type'], trace=aovData['manufactur
er'], response=aovData["co_emissions"],colors=['red', 'blue','green','ye
llow'], ylabel='co_emissions', xlabel='fuel_type',ax=ax)
plt.show()
```

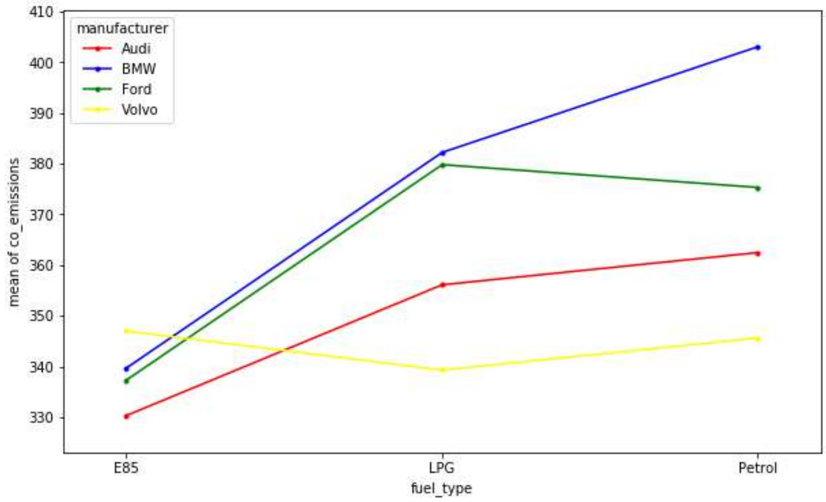

Figure 19: Interaction Plot using fuel_type and manufacturer

The normality assumption on the data has already been tested but homogeneity of variance assumption needs to be checked for both the factors together.

In order to test the assumption that for all 12 groups, the null and alternative hypothesis for population variance can be defined as:

$H_0: \sigma_1 = \sigma_2 \ldots \sigma_{11} = \sigma_{12}$ against H_a: At least one variance is different from the rest.

```
#Aussmption 2: Homogeneity test

statistic, p_value = stats.levene(aovData['co_emissions'][aovData['fuel_type']=="Petrol"][aovData['manufacturer']=="Audi"],
aovData['co_emissions'][aovData['fuel_type']=="E85"][aovData['manufacturer']=="Audi"],aovData['co_emissions'][aovData['fuel_type']=="LPG"][aovData['manufacturer']=="Audi"],aovData['co_emissions'][aovData['fuel_type']=="Petrol"][aovData['manufacturer']=="BMW"],aovData['co_emissions'][aovData['fuel_type']=="E85"][aovData['manufacturer']=="BMW"],aovData['co_emissions'][aovData['fuel_type']=="LPG"][aovData['manufacturer']=="BMW"],
aovData['co_emissions'][aovData['fuel_type']=="Petrol"][aovData['manufacturer']=="Ford"],aovData['co_emissions'][aovData['fuel_type']=="E85"][aovData['manufacturer']=="Ford"],aovData['co_emissions'][aovData['fuel_type']=="LPG"][aovData['manufacturer']=="Ford"],aovData['co_emissions'][aovData['fuel_type']=="Petrol"][aovData['manufacturer']=="Volvo"],aovData['co_emissions'][aovData['fuel_type']=="E85"][aovData['manufacturer']=="Volvo"],aovData['co_emissions'][aovData['fuel_type']=="LPG"][aovData['manufacturer']=="Volvo"])

print("statistic = {}".format(w), "p_value = {}".format(p_value))

statistic = 0.997 p_value = 0.0679
```

Since the p-value is large from 0.05 level of significance therefore, we do not reject the null hypothesis of homogeneity of variances.

Two-way ANOVA procedure is now applied on carbon emission data.

Summary of two-way ANOVA table

```
mod =ols('co_emissions ~ fuel_type*manufacturer', data = aovData).fit()

aov_tbl = sm.stats.anova_lm(mod, type = 2)

print(aov_tbl)
                           df        sum_sq       mean_sq          F    PR(>F)
fuel_type                 2.0  1.028130e+05  51406.481215  12.537086  0.000005
manufacturer              3.0  7.933563e+04  26445.211438   6.449496  0.000273
fuel_type:manufacturer    6.0  5.484639e+04   9141.064215   2.229336  0.039199
Residual                498.0  2.041976e+06   4100.353165        NaN       NaN
```

As noted before (in case of one-way ANOVA), total sum of squares for a given data set is the same. SST for this data is 2278971.

When only fuel type is the predictor, (102813/2278971=) 4.5% of total variability is explained by it. When only manufacturer is the predictor (83825/2278971=) 3.7% of total variability is explained by it. However, when both the factors are in the model (236995/2278971=) 10.4%

of total variability is explained by both main effects and their interaction effects. As more factors are being included in the model, SSE is being reduced. The aim of ANOVA is to explain the total variability in the data, i.e. to assign the variability to definitive causes.

Note also that all three hypotheses are significant at 5% level. Therefore, our conclusion based on two-way ANOVA test, we reject the null hypothesis that all group means are equal for fuel type; we reject the hypothesis that all group means are equal for manufacturers. Similarly, equality of means at each combination of fuel type and manufacturer levels is also rejected.

Residual plots are shown below to see the distribution of residuals at all combinations of fuel type and manufacturer.

```
# model values
model_fitted_y = mod.fittedvalues
# model residuals
model_residuals = mod.resid
# normalized residuals
model_norm_residuals = mod.get_influence().resid_studentized_internal
# absolute squared normalized residuals
model_norm_residuals_abs_sqrt = np.sqrt(np.abs(model_norm_residuals))
# absolute residuals
model_abs_resid = np.abs(model_residuals)
# leverage, from statsmodels internals
model_leverage = mod.get_influence().hat_matrix_diag
# cook's distance, from statsmodels internals
model_cooks = mod.get_influence().cooks_distance[0]

plot_lm_1 = plt.figure()
plot_lm_1.axes[0] = sns.residplot(model_fitted_y, 'co_emissions', data=a
ovData,
                        lowess=True,
                        scatter_kws={'alpha': 0.5},
                        line_kws={'color': 'red', 'lw': 1, 'alpha': 0.8})

plot_lm_1.axes[0].set_title('Residuals vs Fitted')
plot_lm_1.axes[0].set_xlabel('Fitted values')
plot_lm_1.axes[0].set_ylabel('Residuals')
```

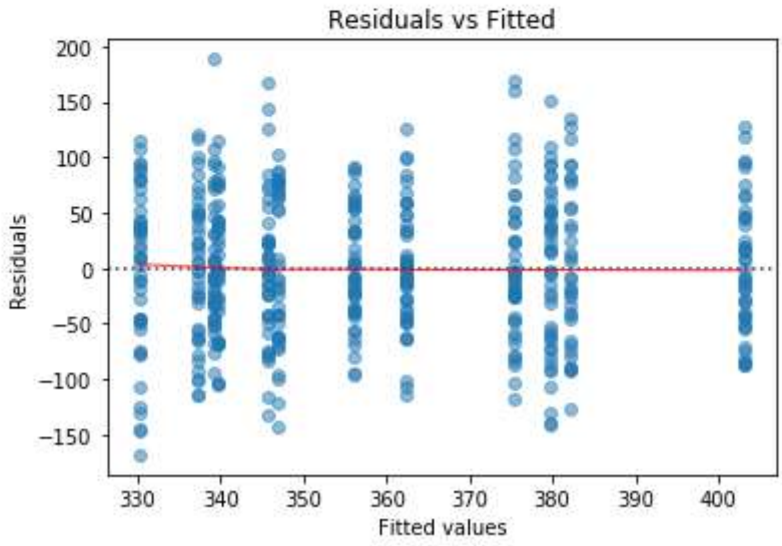

Figure 20: Residuals vs. Fitted plot w.r.t. Fuel Type and manufacturer

Using main Data Frame, means of all factors at each level can be extracted along with their respective counts.

```
# grand means

print('Grand Mean',results.data.mean())

print(np.round(aovData.groupby('fuel_type').agg({'co_emissions':'mean','
Car_ID':'count'}).T,2))



Grand Mean 358.4568823529412

fuel_type         E85     LPG   Petrol

co_emissions   338.12  363.74  371.72

Car_ID         161.00  170.00  179.00

#Car_ID represents value counts



##  manufacturer

print('Grand Mean',results.data.mean())

print(np.round(aovData.groupby('manufacturer').agg({'co_emissions':'mean
','Car_ID':'count'}).T,2))
```

```
Grand Mean 358.4568823529412

manufacturer    Audi     BMW    Ford   Volvo

co_emissions  349.73  377.54  365.08  343.9

Car_ID        142.00  113.00  132.00  123.0


##
##  fuel_type:manufacturer

print(np.round(aovData.groupby(['manufacturer','fuel_type']).agg({'co_em
issions':'mean','Car_ID':'count'}),2))


                        co_emissions  Car_ID

manufacturer fuel_type

Audi         E85              330.33      47

             LPG              356.11      47

             Petrol           362.47      48

BMW          E85              339.70      34

             LPG              382.22      35

             Petrol           403.06      44

Ford         E85              337.32      41

             LPG              379.81      46

             Petrol           375.31      45

Volvo        E85              346.99      39

             LPG              339.30      42

             Petrol           345.62      42
```

Since equality of means hypothesis is rejected, we need to find out which group means are different from the rest.

Tukey-test for multiple comparisons is applied. For the main effects, the results are very similar to one-way ANOVA. Hence these are not included here.

For interaction effect the number of total comparisons are $\binom{12}{2} = 66$. There are 12 combinations and all possible pairs are compared. The results are shown below.

```
###  Posthoc test

MultiComp = MultiComparison(aovData['co_emissions'],aovData['fuel_type']
)
print(MultiComp.tukeyhsd().summary())
results = MultiComp.tukeyhsd()

Multiple Comparison of Means - Tukey HSD, FWER=0.05

====================================================

group1 group2 meandiff p-adj   lower    upper  reject

----------------------------------------------------

   E85    LPG  25.6199 0.0012  8.6837   42.556   True

   E85 Petrol  33.5984  0.001 16.8707 50.3262   True

   LPG Petrol   7.9785 0.4931 -8.5144 24.4715  False

----------------------------------------------------


MultiComp = MultiComparison(aovData['co_emissions'],aovData['manufacture
r'])
print(MultiComp.tukeyhsd().summary())
results = MultiComp.tukeyhsd()


Multiple Comparison of Means - Tukey HSD, FWER=0.05

======================================================

group1 group2 meandiff p-adj   lower    upper   reject

------------------------------------------------------

  Audi    BMW  27.8115 0.0048   6.4089  49.2141   True

  Audi   Ford  15.3513 0.2178  -5.1756  35.8782  False

  Audi  Volvo   -5.829  0.882 -26.7415  15.0835  False

   BMW   Ford -12.4602 0.4541 -34.2191   9.2987  False

   BMW  Volvo -33.6405  0.001 -55.7635 -11.5175   True

  Ford  Volvo -21.1803 0.0516 -42.4573   0.0967  False

------------------------------------------------------



aovData['Car_Fuel'] = aovData.manufacturer + ':' + aovData.fuel_type

MultiComp = MultiComparison(aovData['co_emissions'],aovData['Car_Fuel'])
print(MultiComp.tukeyhsd().summary())
```

```
results = MultiComp.tukeyhsd()
       Multiple Comparison of Means - Tukey HSD, FWER=0.05

==================================================================

   group1      group2    meandiff p-adj    lower    upper   reject

------------------------------------------------------------------

   Audi:E85     Audi:LPG  25.7811 0.6982  -17.5951  69.1572  False

   Audi:E85  Audi:Petrol  32.1377 0.3798   -11.012  75.2874  False

   Audi:E85      BMW:E85   9.3667    0.9  -37.9745  56.7079  False

   Audi:E85      BMW:LPG  51.8848 0.0163    4.9377  98.8319   True

   Audi:E85   BMW:Petrol  72.7291  0.001   28.6198 116.8385   True

   Audi:E85     Ford:E85   6.9898    0.9  -37.9452  51.9249  False

   Audi:E85     Ford:LPG  49.4761 0.0116    5.8648  93.0874   True

   Audi:E85  Ford:Petrol  44.9831 0.0386    1.1277  88.8386   True

   Audi:E85    Volvo:E85  16.6606    0.9  -28.8857  62.2069  False

   Audi:E85    Volvo:LPG   8.9706    0.9  -35.6779   53.619  False

   Audi:E85 Volvo:Petrol  15.2927    0.9  -29.3558  59.9412  False

   Audi:LPG  Audi:Petrol   6.3566    0.9   -36.793  49.5063  False

   Audi:LPG      BMW:E85 -16.4143    0.9  -63.7556  30.9269  False

   Audi:LPG      BMW:LPG  26.1037 0.7767  -20.8434  73.0509  False

   Audi:LPG   BMW:Petrol  46.9481 0.0256    2.8387  91.0574   True

   Audi:LPG     Ford:E85 -18.7912    0.9  -63.7263  26.1439  False

   Audi:LPG     Ford:LPG   23.695 0.8028  -19.9162  67.3063  False

   Audi:LPG  Ford:Petrol  19.2021    0.9  -24.6534  63.0576  False

   Audi:LPG    Volvo:E85  -9.1204    0.9  -54.6667  36.4259  False

   Audi:LPG    Volvo:LPG -16.8105    0.9   -61.459   27.838  False

   Audi:LPG Volvo:Petrol -10.4884    0.9  -55.1368  34.1601  False

Audi:Petrol      BMW:E85   -22.771    0.9  -69.9047  24.3628  False

Audi:Petrol      BMW:LPG   19.7471    0.9  -26.9908   66.485  False

Audi:Petrol   BMW:Petrol  40.5914 0.1011   -3.2952  84.4781  False

Audi:Petrol     Ford:E85 -25.1478 0.7637  -69.8643  19.5686  False
```

Audi:Petrol	Ford:LPG	17.3384	0.9	-26.0476	60.7244	False
Audi:Petrol	Ford:Petrol	12.8454	0.9	-30.786	56.4769	False
Audi:Petrol	Volvo:E85	-15.4771	0.9	-60.8077	29.8536	False
Audi:Petrol	Volvo:LPG	-23.1671	0.8476	-67.5956	21.2613	False
Audi:Petrol	Volvo:Petrol	-16.845	0.9	-61.2734	27.5834	False
BMW:E85	BMW:LPG	42.5181	0.2019	-8.1152	93.1514	False
BMW:E85	BMW:Petrol	63.3624	0.0011	15.3486	111.3763	True
BMW:E85	Ford:E85	-2.3769	0.9	-51.1504	46.3967	False
BMW:E85	Ford:LPG	40.1094	0.1964	-7.4473	87.6661	False
BMW:E85	Ford:Petrol	35.6164	0.3784	-12.1643	83.3972	False
BMW:E85	Volvo:E85	7.2939	0.9	-42.0433	56.6312	False
BMW:E85	Volvo:LPG	-0.3962	0.9	-48.9058	48.1135	False
BMW:E85	Volvo:Petrol	5.926	0.9	-42.5836	54.4356	False
BMW:LPG	BMW:Petrol	20.8443	0.9	-26.781	68.4697	False
BMW:LPG	Ford:E85	-44.8949	0.0985	-93.2861	3.4962	False
BMW:LPG	Ford:LPG	-2.4087	0.9	-49.5731	44.7558	False
BMW:LPG	Ford:Petrol	-6.9017	0.9	-54.292	40.4887	False
BMW:LPG	Volvo:E85	-35.2241	0.4377	-84.1834	13.7351	False
BMW:LPG	Volvo:LPG	-42.9142	0.1338	-91.0394	5.2109	False
BMW:LPG	Volvo:Petrol	-36.5921	0.3451	-84.7172	11.533	False
BMW:Petrol	Ford:E85	-65.7393	0.001	-111.3825	-20.0961	True
BMW:Petrol	Ford:LPG	-23.253	0.8415	-67.5936	21.0875	False
BMW:Petrol	Ford:Petrol	-27.746	0.6408	-72.3268	16.8348	False
BMW:Petrol	Volvo:E85	-56.0685	0.0045	-102.3136	-9.8234	True
BMW:Petrol	Volvo:LPG	-63.7586	0.001	-109.1197	-18.3975	True
BMW:Petrol	Volvo:Petrol	-57.4364	0.0022	-102.7975	-12.0754	True
Ford:E85	Ford:LPG	42.4863	0.0875	-2.6758	87.6483	False
Ford:E85	Ford:Petrol	37.9933	0.2062	-7.4047	83.3912	False
Ford:E85	Volvo:E85	9.6708	0.9	-37.3626	56.7041	False
Ford:E85	Volvo:LPG	1.9807	0.9	-44.1837	48.1451	False

```
   Ford:E85 Volvo:Petrol   8.3028    0.9  -37.8616  54.4673  False
   Ford:LPG  Ford:Petrol   -4.493    0.9   -48.581  39.5951  False
   Ford:LPG    Volvo:E85 -32.8155 0.4434  -78.5857  12.9548  False
   Ford:LPG    Volvo:LPG -40.5055 0.1232  -85.3825   4.3714  False
   Ford:LPG Volvo:Petrol -34.1834 0.3422  -79.0603  10.6935  False
Ford:Petrol    Volvo:E85 -28.3225 0.6545  -74.3255  17.6805  False
Ford:Petrol    Volvo:LPG -36.0126 0.2708  -81.1269   9.1017  False
Ford:Petrol Volvo:Petrol -29.6904 0.5675  -74.8047  15.4238  False
  Volvo:E85    Volvo:LPG  -7.6901    0.9  -54.4497  39.0695  False
  Volvo:E85 Volvo:Petrol  -1.3679    0.9  -48.1275  45.3916  False
  Volvo:LPG Volvo:Petrol   6.3221    0.9  -39.5634  52.2076  False
------------------------------------------------------------------
```

Below is shown the plot for quick visual comparison.

```
df=results.summary()
results_as_html = df.as_html()
df1=pd.read_html(results_as_html, header=0, index_col=0)[0].reset_index(
)
groups = np.array([df1.group1+ '-'+ df1.group2])

plt.figure(figsize=(10,20))
data_dict = {}
data_dict['category'] = groups.ravel()
data_dict['lower'] = results.confint[:,0]
data_dict['upper'] = results.confint[:,1]
dataset = pd.DataFrame(data_dict)

for lower,upper,y in zip(dataset['lower'],dataset['upper'],range(len(dat
aset))):
    plt.plot((lower,upper),(y,y),'ro-',color='orange')
plt.yticks(range(len(dataset)),list(dataset['category']));
```

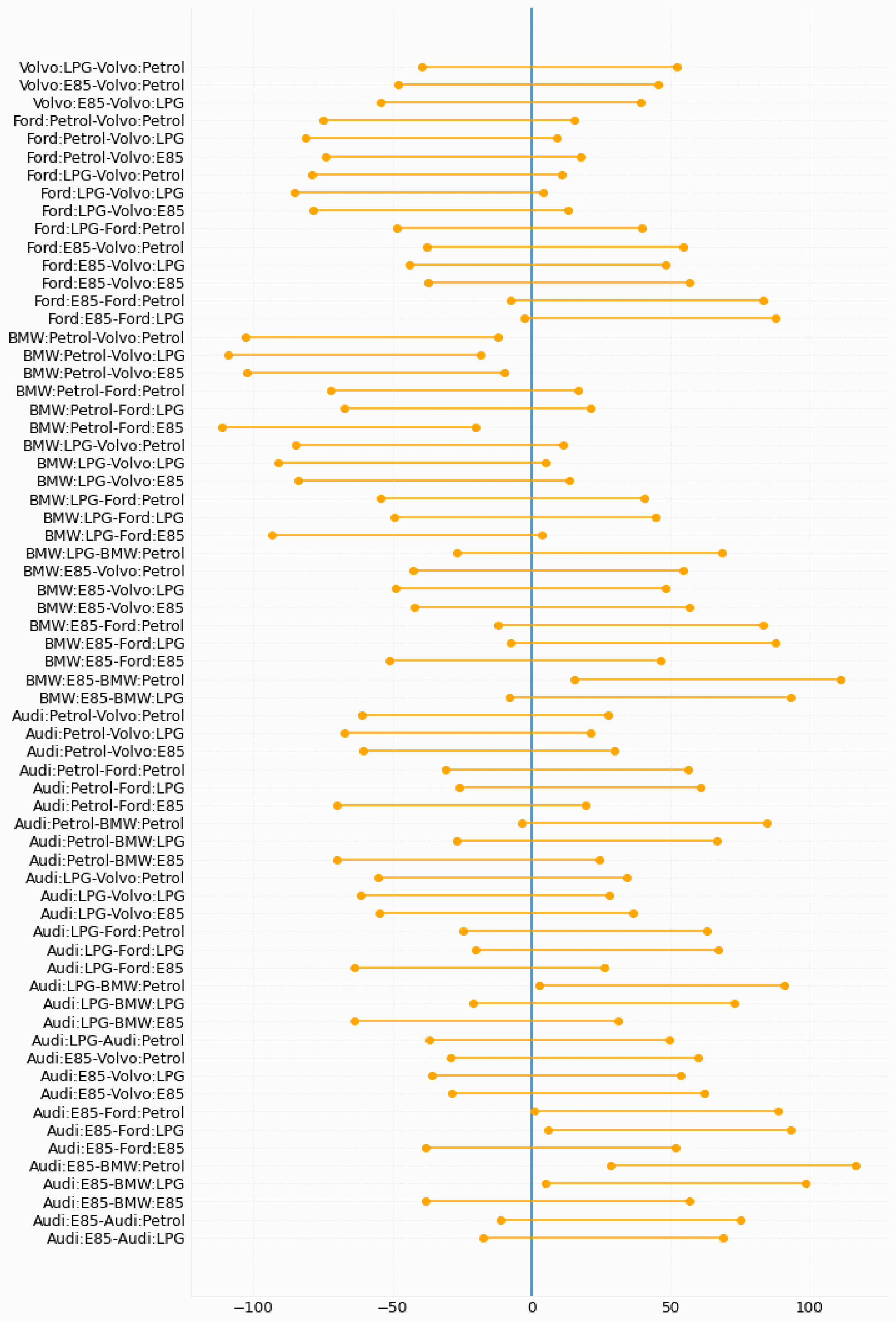

Figure 21: Family-wise comparison for Fuel type and Manufacturer

Significant means are highlighted with grey color or can be seen in figure 22. For example, the BMW car having fuel type (LPG and Petrol) and Ford manufacturer with (LPG and Petrol) has significantly different mean from Audi manufacturer with E85 fuel type. Similarly, other significant difference can also be observed between BMW (having petrol) & Audi (with LPG), Ford (with E85) & BMW (with Petrol). Also, different fuel type (Petrol and E85) though are from the same manufacturer has a significant mean difference.

Conclusion: It is observed that the variation in releasing Carbon Emission is significantly impacted by fuel type and manufacturers along with their interaction effect. ANOVA helps in identifying which independent factor(s) can explain the variation in the response variable.

Summary Chart for ANOVA

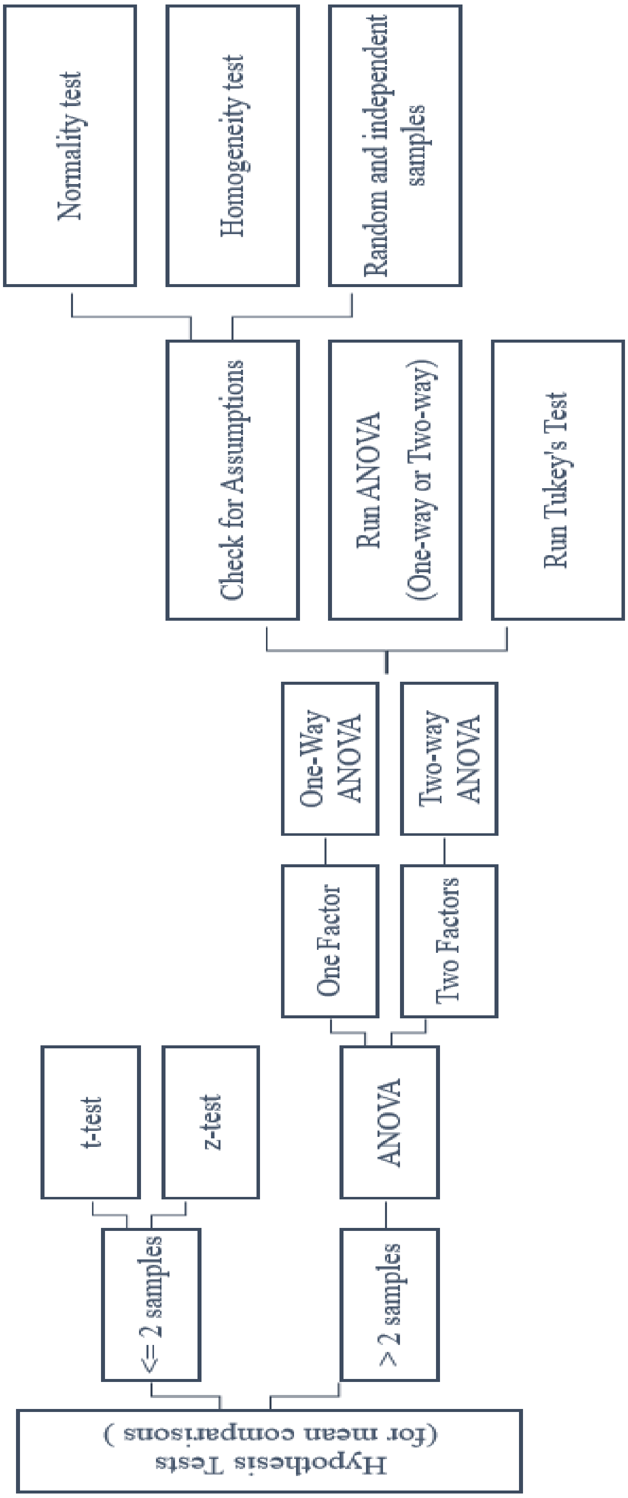

Figure 22: Summary of ANOVA

5. References:

Analysis of Variance and design of Experiments, PennState Eberly College of Science, https://newonlinecourses.science.psu.edu/stat502/node/137/

Black, K. (2011). Applied business statistics: Making better business decisions. Wiley.

Kothari, C. R. (2004). Research methodology: Methods and techniques. New Age International.

Malhotra, N., Hall, J., Shaw, M., & Oppenheim, P. (2006). Marketing research: An applied orientation. Pearson Education Australia.

McHugh, M. L. (2011). Multiple comparison analysis testing in ANOVA. Biochemia medica: Biochemia medica, 21(3), 203-209.

Rutherford, A. (2001). Introducing ANOVA and ANCOVA: a GLM approach. Sage.

Greenwood, M. and Banner, K., “2.5 - Multiple (pair-wise) comparisons using Tukey's HSD and the compact letter display”, Accessed date: 02 April’ 2019. https://arc.lib.montana.edu/book/statistics-with-r-textbook/item/59#2.5%20-%20Multiple%20(pair-wise)%20comparisons%20using%20Tukey's%20HSD%20and%20the%20compact%20letter%20display