Question 1 (Data Analysis)

The analysis was done by using SAS Enterprise Guide.

i. Statistical Analysis

a) Descriptive Analysis

Summary Statistics

Results

The MEANS Procedure

Variable	Mean	Std Dev	Minimum	Maximum	N
a	1.5183654	0.0030369	1.5111500	1.5339300	214
b	13.4078505	0.8166036	10.7300000	17.3800000	214
c	2.6845327	1.4424078	0	4.4900000	214
d	1.4449065	0.4992696	0.2900000	3.5000000	214
e	72.6509346	0.7745458	69.8100000	75.4100000	214
f	0.4970561	0.6521918	0	6.2100000	214
g	8.9569626	1.4231535	5.4300000	16.1900000	214
h	0.1750467	0.4972193	0	3.1500000	214
i	0.0570093	0.0974387	0	0.5100000	214

Correlation

Pearson Correlation Coefficients, N = 214 Prob > |r| under H0: Rho=0	
	a
b	-0.19189
	0.0049
c	-0.12227
	0.0743
d	-0.40733
	<.0001
e	-0.54205
	<.0001
f	-0.28983
	<.0001
g	0.81040
	<.0001
h	-0.00039
	0.9955
i	0.14301
	0.0366

The first correlation test is between additive “a” with the other eight additives. From the result above, the interpretation of the test can be seen in the table below:

Relationship between additive “a” with additive:	**Detail**
b	At level of significance of 5%, there are weak and negative relationship between additive “a” and additive “b”
c	As p-value > 5%, there is no relationship between the additives.
d	At level of significance of 5%, there are below average and negative relationship between additive “a” and additive “d”
e	At level of significance of 5%, there are above average and negative relationship between additive “a” and additive “e”
f	At level of significance of 5%, there are below average and negative relationship between additive “a” and additive “f”
g	At level of significance of 5%, there are strong and positive relationship between additive “a” and additive “g”
h	As p-value > 5%, there is no relationship between the additives.
i	At level of significance of 5%, there are weak and positive relationship between additive “a” and additive “i”

Pearson Correlation Coefficients, N = 214 Prob > |r| under H0: Rho=0	
	b
c	-0.27373
	<.0001
d	0.15679
	0.0218
e	-0.06981
	0.3094
f	-0.26609
	<.0001
g	-0.27544
	<.0001
h	0.32660
	<.0001
i	-0.24135
	0.0004

The second correlation test is between additive “b” with the other seven additives (excluding additive ‘a’ because it had been shown in the first test). From the result above, the interpretation of the test can be seen in the table below:

Relationship between additive “b” with additive:	**Detail**
c	At level of significance of 5%, there are below average and negative relationship between additive “b” and additive “c”
d	At level of significance of 5%, there are weak and positive relationship between additive “b” and additive “d”
e	As p-value > 5%, there is no relationship between the additives.
f	At level of significance of 5%, there are below average and negative relationship between additive “b” and additive “f”
g	At level of significance of 5%, there are below average and negative relationship between additive “b” and additive “g”
h	At level of significance of 5%, there are below average and positive relationship between additive “b” and additive “h”
i	At level of significance of 5%, there are below average and negative relationship between additive “b” and additive “i”

Pearson Correlation Coefficients, N = 214 Prob > |r| under H0: Rho=0	
	c
d	-0.48180
	<.0001
e	-0.16593
	0.0151
f	0.00540
	0.9375
g	-0.44375
	<.0001
h	-0.49226
	<.0001
i	0.08306
	0.2263

The third correlation test is between additive “c” with the other six additives (excluding additive ‘a’ and additive ‘b’ because they had been shown in the previous tests). From the result above, the interpretation of the test can be seen in the table below:

Relationship between additive “c” with additive:	**Detail**
d	At level of significance of 5%, there are below average and negative relationship between additive “c” and additive “d”
e	At level of significance of 5%, there are weak and negative relationship between additive “c” and additive “e”
f	As p-value > 5%, there is no relationship between the additives.
g	At level of significance of 5%, there are below average and negative relationship between additive “c” and additive “g”
h	At level of significance of 5%, there are below average and negative relationship between additive “c” and additive “h”
i	As p-value > 5%, there is no relationship between the additives.

Pearson Correlation Coefficients, N = 214 Prob > |r| under H0: Rho=0	
	d
e	-0.00552
	0.9360
f	0.32596
	<.0001
g	-0.25959
	0.0001
h	0.47940
	<.0001
i	-0.07440
	0.2786

The fourth correlation test is between additive “d” with the other five additives (excluding additive ‘a’, additive ‘b’, and additive ‘c’ because they had been shown in the previous tests). From the result above, the interpretation of the test can be seen in the table below:

Relationship between additive "d" with additive:	Detail
e	As p-value > 5%, there is no relationship between the additives.
f	At level of significance of 5%, there are below average and positive relationship between additive "d" and additive "f"
g	At level of significance of 5%, there are below average and negative relationship between additive "d" and additive "g"
h	At level of significance of 5%, there are below average and positive relationship between additive "d" and additive "h"
i	As p-value > 5%, there is no relationship between the additives.

Pearson Correlation Coefficients, N = 214 Prob > |r| under H0: Rho=0	
	e
f	-0.19333
	0.0045
g	-0.20873
	0.0021
h	-0.10215
	0.1364
i	-0.09420
	0.1697

The fifth correlation test is between additive "e" with the other four additives (excluding additive 'a', additive 'b', additive 'c', and additive 'd' because they had been shown in the previous tests). From the result above, the interpretation of the test can be seen in the table below:

Relationship between additive "e" with additive:	Detail
f	At level of significance of 5%, there are weak and negative relationship between additive "e" and additive "f"
g	At level of significance of 5%, there are below average and negative relationship between additive "e" and additive "g"

h	As p-value > 5%, there is no relationship between the additives.
i	As p-value > 5%, there is no relationship between the additives.

Pearson Correlation Coefficients, N = 214 Prob > |r| under H0: Rho=0	
	f
g	-0.31784
	<.0001
h	-0.04262
	0.5352
i	-0.00772
	0.9106

The sixth correlation test is between additive "f" with the other three additives (excluding additive 'a', additive 'b', additive 'c', additive 'd' and additive 'e' because they had been shown in the previous tests). From the result above, the interpretation of the test can be seen in the table below:

Relationship between additive "f" with additive:	**Detail**
g	At level of significance of 5%, there are below average and negative relationship between additive "f" and additive "g"
h	As p-value > 5%, there is no relationship between the additives.
i	As p-value > 5%, there is no relationship between the additives.

Pearson Correlation Coefficients, N = 214 Prob > |r| under H0: Rho=0	
	g
h	-0.11284
	0.0997
i	0.12497
	0.0681

The seventh correlation test is between additive “g” with the other two additives (excluding additive ‘a’, additive ‘b’, additive ‘c’, additive ‘d’, additive ‘e’, and additive ‘f’ because they had been shown in the previous tests). From the result above, the interpretation of the test can be seen in the table below:

Relationship between additive “g” with additive:	**Detail**
h	As p-value > 5%, there is no relationship between the additives.
i	As p-value > 5%, there is no relationship between the additives.

Pearson Correlation Coefficients, N = 214 Prob > |r| under H0: Rho=0	
	h
i	-0.05869 0.3929

The last correlation test is between additive “h” with the one remaining additive (excluding additive ‘a’, additive ‘b’, additive ‘c’, additive ‘d’, additive ‘e’, additive ‘f’, and additive ‘g’ because they had been shown in the previous tests). From the result above, the interpretation of the test can be seen in the table below:

Relationship between additive “h” with additive:	**Detail**
i	As p-value > 5%, there is no relationship between the additives.

ANOVA

The ANOVA Procedure

Dependent Variable: score

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	943261.0844	117907.6356	168332	<.0001
Error	1917	1342.7570	0.7004		
Corrected Total	1925	944603.8415			

R-Square	Coeff Var	Root MSE	score Mean
0.998578	7.428885	0.836927	11.26585

Source	DF	Anova SS	Mean Square	F Value	Pr > F
additives	8	943261.0844	117907.6356	168332	<.0001

The ANOVA F-Test procedure can be described in the steps below:

Step 1:

H_0: All the means are the same

H_1: At least one mean is different from others

Step 2:

p-value < level of significance

0.0001 < 0.05

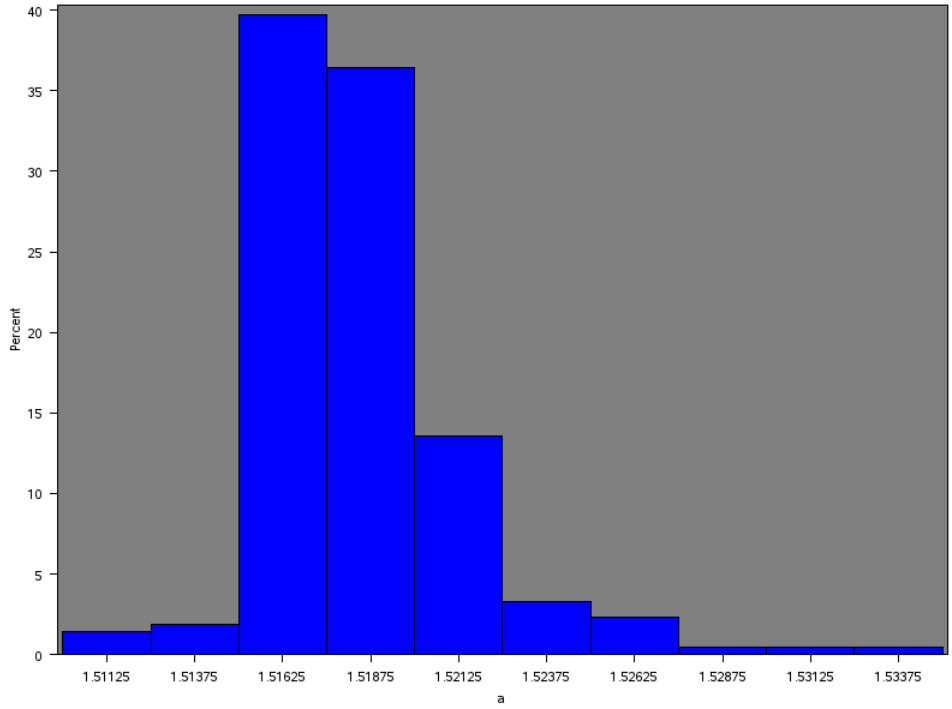
Step 3:

Since the p-value is smaller than level of significance, hence H_0 is rejected.

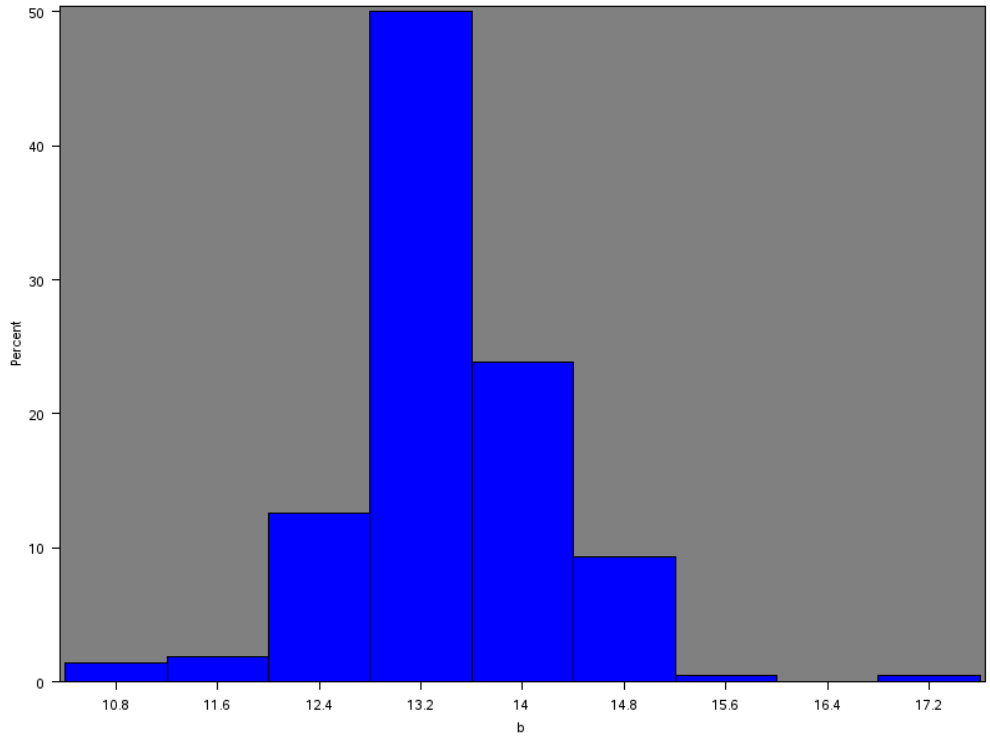
Step 4:

At level of significance of 5%, it can be concluded that at least one mean is different from others, which means the formulations are not equally effective.

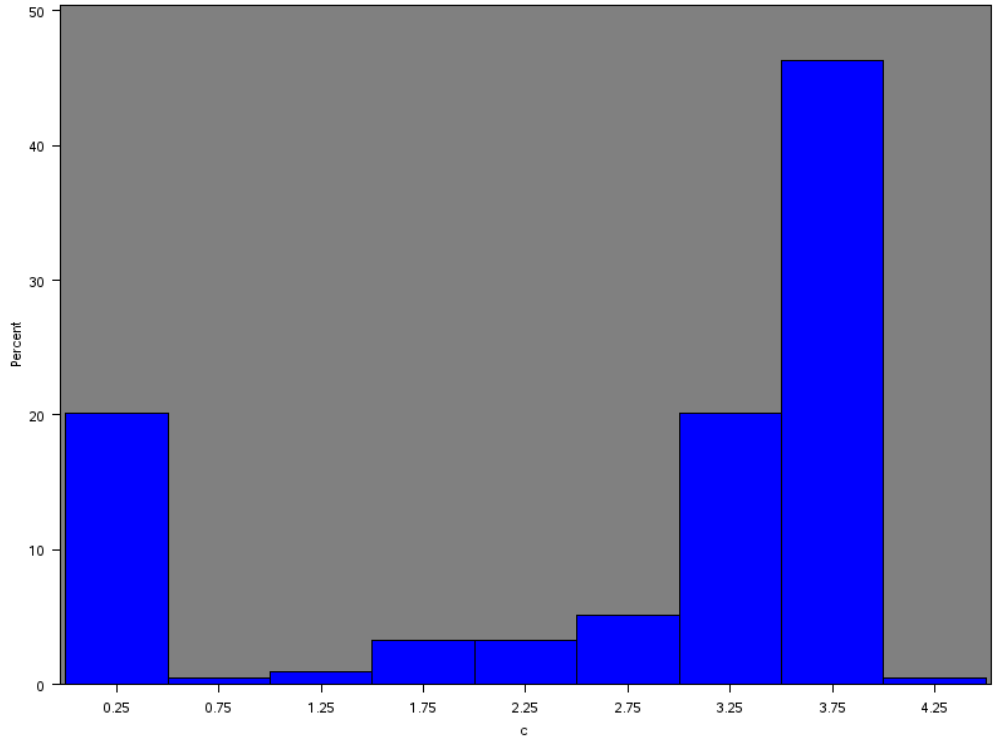
b) Graphical Analysis



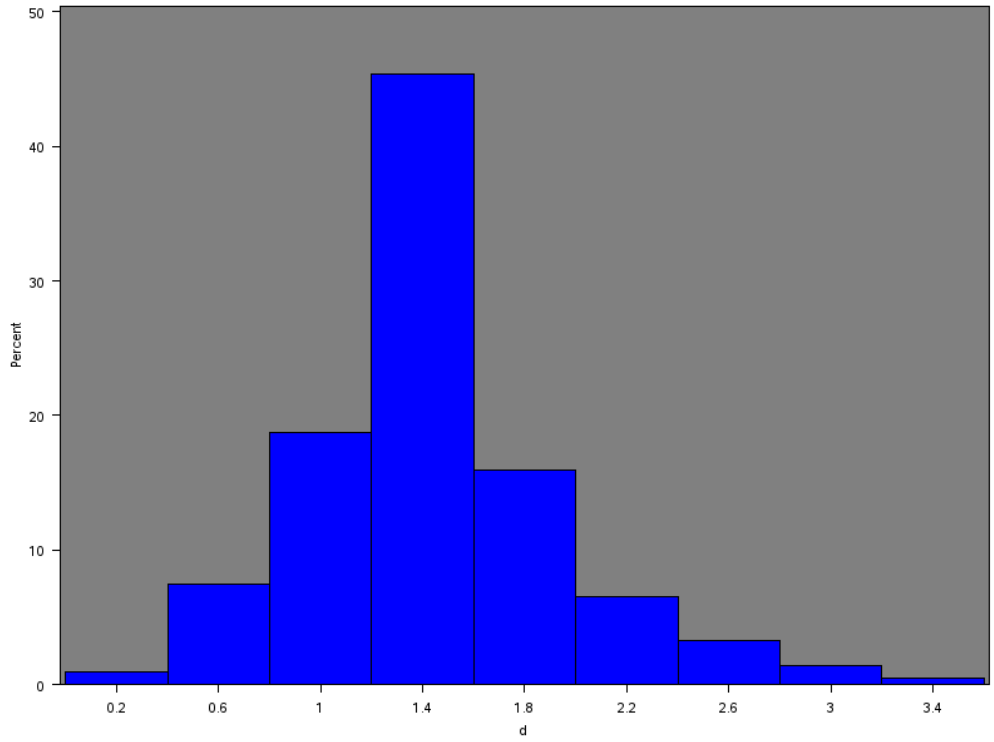
The figure above shows distribution chart for additive ‘a’. It can be seen that the data is distributed mostly in the range of 1.51625 to 1.51875, and is skewed to the left.



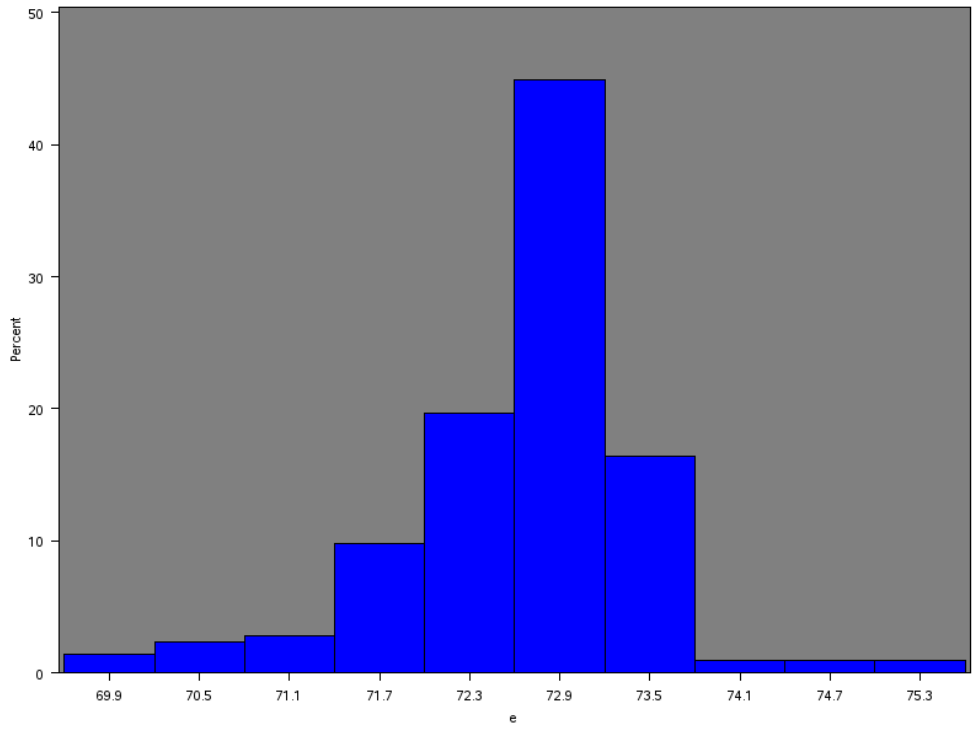
The figure above shows distribution chart for additive ‘b’. It can be seen that the data is distributed mostly around 13.2.



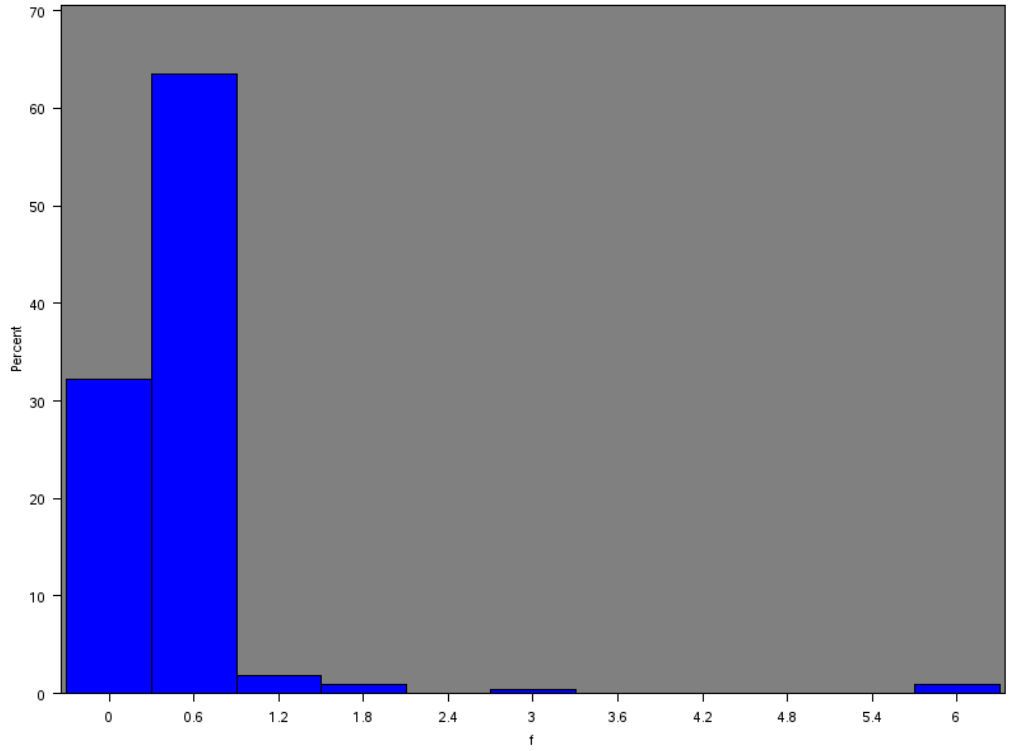
The figure above shows distribution chart for additive ‘c’. It can be seen that the data is heavily distributed around 0.25 and in the range of 3.25 to 3.75.



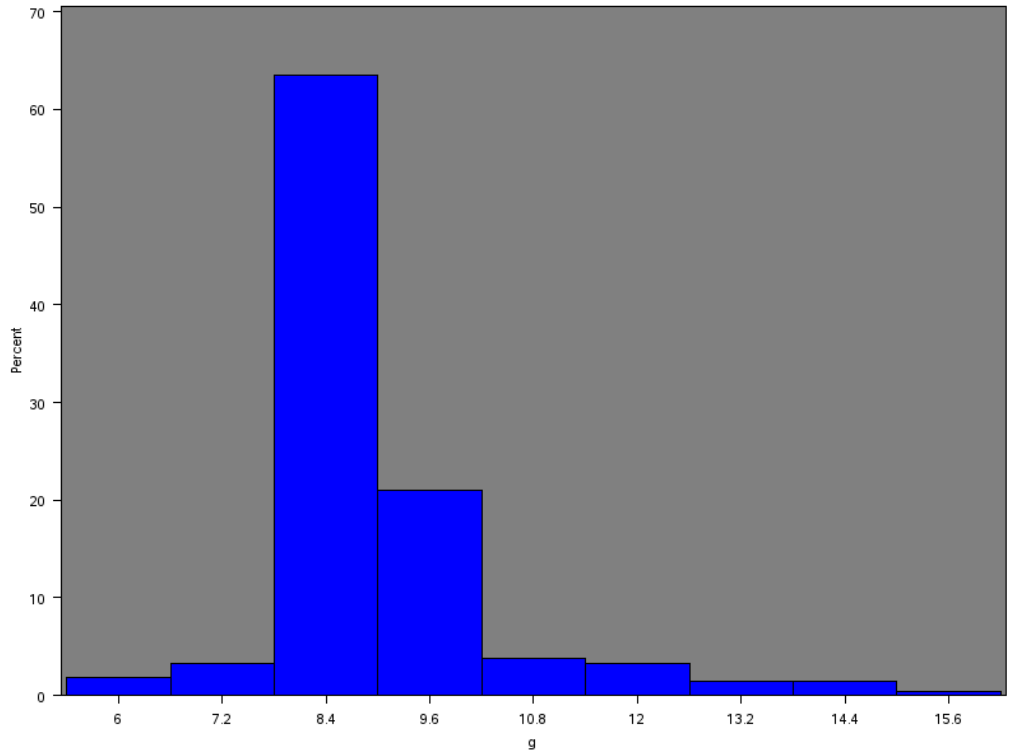
The figure above shows distribution chart for additive ‘d’. It can be seen that the data is distributed mostly around 1.4.



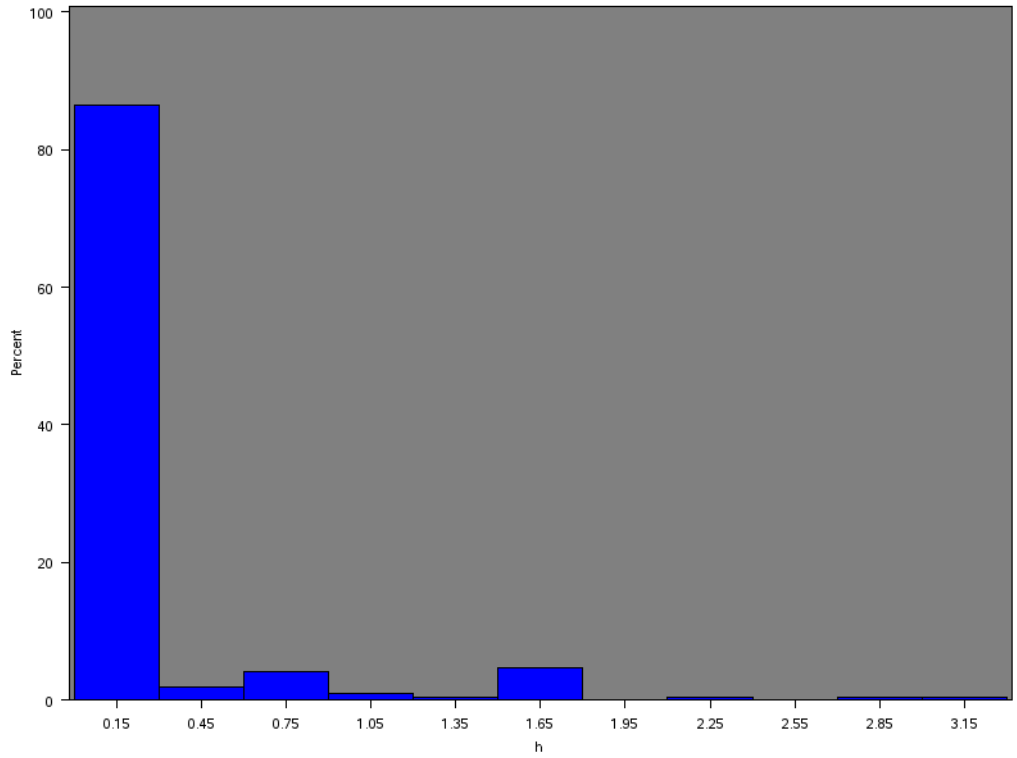
The figure above shows distribution chart for additive ‘e’. It can be seen that the data is distributed mostly around 72.9.



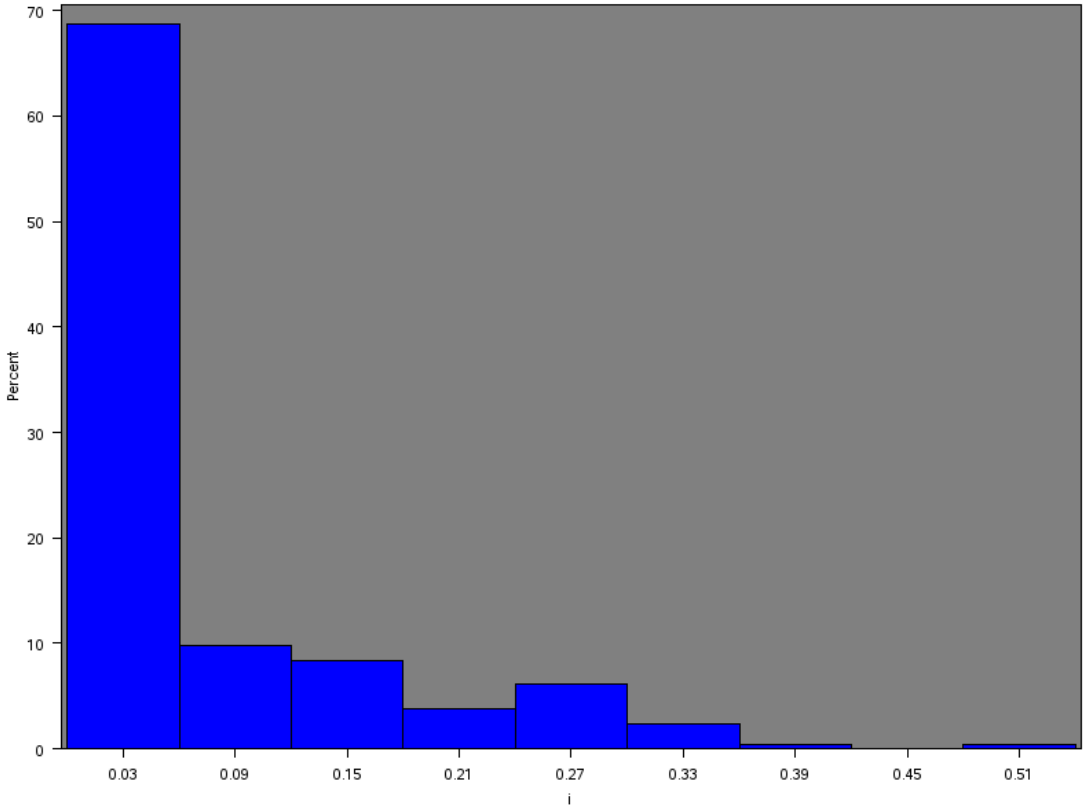
The figure above shows distribution chart for additive ‘f’. It can be seen that the data is distributed mostly in the range of 0 to 0.6. There are outliers with the value 3 and 6.



The figure above shows distribution chart for additive ‘g’. It can be seen that the data is distributed mostly around 8.4.



The figure above shows distribution chart for additive ‘h’. It can be seen that the data is distributed mostly around 0.15.



The figure above shows distribution chart for additive ‘i’. It can be seen that the data is distributed mostly around 0.03.

c) Clustering (k-means)

SAS Enterprise Guide is used to perform the k-means clustering. Part of the output is shown as below.

Cluster Standard Deviations									
Cluster	a	b	c	d	e	f	g	h	i
1	0.001945060	1.166088255	0.719058728	0.503007320	0.763705083	0.782446048	0.794607965	0.000000000	0.085450677
2	.	.	.	.	.	.	.	.	.
3	0.001368145	0.369044742	0.496935537	0.370780610	0.373810917	0.286290620	0.605961776	0.640352598	0.032142685
4	0.003539101	0.315013227	0.318956632	0.881532756	1.461141107	0.155349069	0.234378611	0.506392470	0.000000000
5	0.002213918	0.469344285	0.931496770	0.496004142	0.503824324	0.215445275	0.569025196	0.072129865	0.119393850
6	0.001566587	0.444527383	0.251714516	0.255799310	0.456349696	0.169631069	0.435536442	0.063423427	0.098875819
7	0.002484398	0.503620227	0.736636500	0.376342043	1.247089946	0.382143080	0.725051722	0.969948452	0.000000000
8	0.000035355	0.014142136	0.000000000	0.014142136	0.155563492	0.000000000	0.021213203	0.000000000	0.000000000
9	0.003879693	1.067386662	0.000000000	0.443393972	0.980086549	0.268031848	1.339700793	0.000000000	0.133844259
10	.	.	.	.	.	.	.	.	.

It can be seen that eight distinctive clusters are formed with members as follow:

Cluster 1: b, c, e, f, g

Cluster 3: g, h

Cluster 4: d, e

Cluster 5: c

Cluster 6: b, e, g

Cluster 7: e, h

Cluster 8: b, d, e, g

Cluster 9: b, e, g

ii. Text Analysis

The analysis was performed by using Python.
Kindly refer to the Python file named “Q1_TextAnalysis.py”.

a) Probability of word [data] occuring in line 1 is 0.034482758620689655
Probability of word [data] occuring in line 2 is 0.05555555555555555
Probability of word [data] occuring in line 3 is 0.037037037037037035
Probability of word [data] occuring in line 4 is 0.024390243902439025
Probability of word [data] occuring in line 5 is 0.037037037037037035
Probability of word [data] occuring in line 6 is 0.06666666666666667
Probability of word [data] occuring in line 7 is 0.11904761904761904
Probability of word [data] occuring in line 8 is 0.02564102564102564
Probability of word [data] occuring in line 9 is 0.21428571428571427
Probability of word [data] occuring in line 10 is 0.058823529411764705
Probability of word [data] occuring in line 11 is 0.03333333333333333

b) The distribution of distinct word counts across the lines is as follows:
Word (space) Count
as 2
a 10
term 3
data 18
analytics 10
predominantly 1
refers 1
to 11
an 1
assortment 1
of 10
applications 1
from 2
basic 1
business 4

intelligence 1
bi 2
reporting 1
and 9
online 1
analytical 1
processing 1
olap 1
various 1
forms 1
advanced 2
in 6
that 4
sense 1
it's 1
similar 1
nature 1
another 1
umbrella 1
for 2
approaches 1
analyzing 1
with 3
the 11
difference 1
latter 1
is 4
oriented 1
uses 2
while 2
has 2
broader 1
focus 1
expansive 1

view 2
isn't 1
universal 1
though 1
some 1
cases 1
people 1
use 1
specifically 1
mean 1
treating 1
separate 1
category 1
initiatives 1
can 5
help 1
businesses 1
increase 1
revenues 1
improve 1
operational 1
efficiency 1
optimize 1
marketing 1
campaigns 1
customer 1
service 1
efforts 1
respond 1
more 2
quickly 1
emerging 1
market 1
trends 1

gain 1
competitive 1
edge 1
over 1
rivals 1
all 1
ultimate 1
goal 1
boosting 1
performance 1
depending 1
on 2
particular 1
application 1
that's 1
analyzed 1
consist 1
either 1
historical 1
records 1
or 4
new 1
information 1
been 1
processed 1
real 1
time 1
addition 1
it 2
come 1
mix 1
internal 1
systems 1
external 1

sources 1
at 1
high 1
level 1
methodologies 1
include 1
exploratory 2
analysis 6
eda 2
which 2
aims 1
find 1
patterns 1
relationships 1
confirmatory 1
cda 2
applies 1
statistical 1
techniques 1
determine 1
whether 1
hypotheses 1
about 1
set 1
are 1
true 1
false 1
often 1
compared 2
detective 1
work 2
akin 1
judge 1
jury 1

during 1
court 1
trial 1
distinction 1
first 1
drawn 1
by 1
statistician 1
john 1
w 1
tukey 1
his 1
1977 1
book 1
also 1
be 2
separated 1
into 1
quantitative 1
qualitative 2
former 1
involves 1
numerical 2
quantifiable 1
variables 1
measured 1
statistically 1
approach 1
interpretive 1
focuses 1
understanding 1
content 1
non 1
like 1

text 1

images 1

audio 1

video 1

including 1

common 1

phrases 1

themes 1

points 1

c) The probability of the word [analytics] occuring after the word [data] is:
0.3333333333333333

2. Machine Learning (Linear Regression)

The question was answered by using Python.
Kindly refer to the Python file named “Q2_LinearRegression.py”.

3. Web Scraping

The question was answered by using Python.
Kindly refer to the Python file named “Q3_WebScraping.py”

For this question, I was unable to perform web scraping to the target website effectively due to problem in handling drop-down list in the website. Thus, the list of the jobs and skills are incomplete.

4. Bonus Points

Kindly refer to “Capstone Project.pdf” and “SpellChecking.zip” files for my previous projects as part of my postgraduate study modules.