STATISTICAL NATURAL LANGUAGE PROCESSING
ASSIGNMENT

COURSE NAME
INTRODUCTION TO DATA SCIENCE WITH PYTHON

COURSE CODE PM-ASDS04

SUBMITTED TO

DR. AJIT KUMAR MAJUMDER

PROFESSOR, DEPARTMENT OF STATISTICS
JAHANGIRNAGOR UNIVARSITY

SUMITTED FROM

MD JAHIDUL ALAM

PM-ASDS-BATCH A
ROLL-201900101051

Introduction: Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. We will discuss about some statistical calculation which considered a major part in Data Science platform. The discussion is related with Random Number Generation and Regression Analysis from data.

Objectives:

Objectives of Random Number Generation: To generate a sequence of numbers or symbols that cannot be reasonably predicted well than by a random chance.

Objectives of Regression Analysis: To understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables.

Methodology:

Random Number Generation: A random number generator (RNG) is a mathematical construct, either computational or as a hardware device, that is designed to generate a random set of numbers that should not display any distinguishable patterns in their appearance or generation, hence the word random.

Random Number Generation working procedure in excel:

I. To generate random numbers, first click the Data tab's Data Analysis command button.
II. In the Data Analysis dialog box, select the Random Number Generation entry from the list and then click OK. Excel displays the Random Number Generation dialog box.
III. Then we describe how many columns and rows of values that we want.
IV. Select the distribution method.
V. (Optional) Provide any parameters needed for the distribution method.
VI. (Optional) Select a starting point for the random number generation.
VII. Identify the output range.
VIII. After we describe how we want Excel to generate random numbers and where those numbers should be placed, click OK. Excel generates the random numbers.

Regression Analysis: Regression Analysis includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. A regression model is a mathematical equation that describes the relationship between two or more variables. A regression model that includes two or more independent variables is called a multiple regression model. It is written as

$$y = A + B1x1 + B2x2 + B3x3+ \ldots + Bkxk + \varepsilon$$

where y is the dependent variable, x1, x2, x3, …, xk are the k independent variables, and ε is the random error term.

Regression Analysis working procedure in excel:

I. On the Data tab, in the Analysis group, click the Data Analysis button.
II. Select Regression and click OK.

III. In the Regression dialog box, configure the following settings: Select the Input Y Range, which is your dependent variable. Select the Input X Range, i.e. our independent variable. If we are building a multiple regression model, select two or more adjacent columns with different independent variables.

- Check the Labels box if there are headers at the top of your X and Y ranges.
- Choose your preferred Output option, a new worksheet in our case.
- Optionally, select the Residuals checkbox to get the difference between the predicted and actual values. IV. Click OK and observe the regression analysis output created by Excel.

Results Analysis:

1. We generate unique random numbers from the given data (which data is provided in our classroom) in Figure-01.

	Item Name	Calories	Protein (g)	Total Fat (g)	
2	Apple Slices	15	0	0	1.432050539
3	Bacon, Egg & Cheese Bagel	630	30	32	73.47572253
4	Bacon, Egg & Cheese Bagel with Egg Whites	580	30	26	10.26642659
5	Bacon, Egg & Cheese Biscuit (Large Size Biscuit)	520	19	30	57.7980285
6	Bacon, Egg & Cheese Biscuit (Regular Size Biscuit)	460	19	26	34.96279794
7	Bacon, Egg & Cheese Biscuit with Egg Whites (Large Biscuit)	470	20	25	20.27610096
8	Bacon, Egg & Cheese Biscuit with Egg Whites (RegularBiscuit)	410	20	20	30.18607135
9	Bacon, Egg & Cheese McGriddles	460	19	21	13.53248695
10	Bacon, Egg & Cheese McGriddles with Egg Whites	400	20	15	44.90902432
11	Baked Hot Apple Pie	250	2	13	93.09625538
12	Big Breakfast with Egg Whites (Large Size Biscuit)	690	26	41	4.383892331
13	Big Breakfast with Egg Whites (Regular Size Biscuit)	640	26	37	90.15952025
14	Big Breakfast with Hotcakes (Large Size Biscuit)	1150	36	60	65.83175146
15	Big Breakfast with Hotcakes (Regular Size Biscuit)	1090	36	56	17.5417951
16	Big Breakfast with Hotcakes and Egg Whites (Large Biscuit)	1050	35	50	30.57884457
17	Big Breakfast with Hotcakes and Egg Whites (Regular Biscuit)	990	35	46	42.34391308
18	Big Breakfast® (Large Size Biscuit)	800	28	52	46.7792291
19	Big Breakfast® (Regular Size Biscuit)	740	28	48	79.760094
20	Big Mac	550	25	29	88.05063021
21	Blueberry Pomegranate Smoothie (Large)	340	4	1	42.70949431
22	Blueberry Pomegranate Smoothie (Medium)	260	3	1	79.02288888
23	Blueberry Pomegranate Smoothie (Small)	220	2	1	10.04284799

2. We also sort the generated random numbers from smallest to largest in Figure-02.

1	Item Name	Calories	Protein (g)	Total Fat (g)	
2	McCafé Caramel Hot Chocolate with Nonfat Milk (Large)	380	18	4	1.265877255
3	Apple Slices	15	0	0	1.432050539
4	Chicken McNuggets® (10 piece)	470	22	30	1.519669179
5	Quarter Pounder Bacon & Cheese	600	37	29	1.858058412
6	Kids Fries	100	1	5	2.042359691
7	Premium Grilled Chicken Club Sandwich	510	40	20	2.317300943
8	Iced Coffee– Regular (Medium)	190	1	7	2.531815546
9	Mocha (Small)	340	10	11	2.580156865
10	Premium Southwest Salad (without chicken)	140	6	5	2.921567431
11	Big Breakfast with Egg Whites (Large Size Biscuit)	690	26	41	4.383892331
12	Mango Pineapple Smoothie (Large)	340	5	2	5.465529344
13	Sausage McMuffin with Egg Whites	400	22	22	6.205755791
14	Nonfat Latte with Sugar Free French Vanilla Syrup (Large)	220	16	1	6.474654378
15	Cheeseburger	300	15	12	6.498825037
16	Hot Caramel Sundae	340	7	8	6.504867702
17	Sausage Biscuit (Large Size Biscuit)	480	11	31	6.668019654
18	Daily Double	440	23	24	9.036744285
19	Premium McWrap Chicken Sweet Chili (Crispy)	520	23	22	9.595690786
20	Hazelnut Latte (Medium)	330	11	10	9.637989441
21	Nonfat Caramel Latte (Small)	200	10	0	9.81020539
22	Hotcakes	350	8	9	9.879696036
23	Blueberry Pomegranate Smoothie (Small)	220	2	1	10.04284799

3. Regression Analysis: Calories versus Protein, Total Fat from given data (which data is provided in our classroom) in Figure-03.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.687950527
R Square	0.473275928
Adjusted R Squ	0.468573034
Standard Error	149.7469774
Observations	227

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	4513312.42	2256656.21	100.6350512	6.56621E-32
Residual	224	5023011.219	22424.15723		
Total	226	9536323.639			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	177.1986095	17.88699065	9.906563544	2.02734E-19	141.9503096	212.447	141.9503	212.4469093
0	13.88397586	1.042313392	13.3203468	3.34586E-30	11.82998166	15.938	11.82998	15.93797005
0	0.899207923	0.282934669	3.178146839	0.001691349	0.341653764	1.45676	0.341654	1.456762083

Discussion:

Let, y = Calories x1 = Portion x2 = Total Fat

We are to estimate the regression model

$y = A + B1x1 + B2x2 + \varepsilon$

From the output given in Screen, the estimated regression equation is:

$y = 177.19 + 13.88x1 + 0.89x2$

The value of a = 177.19 in the estimated regression equation gives the value of y for x1 = 0 and x2 = 0. The value of b1 = 13.88 in the estimated regression model gives the change in y for a one-unit change in x1 when x2 is held constant.

The value of b2 = 0.89 in the estimated regression model gives the change in y for a one-unit change in x2 when x1 is held constant. Here P-value is approximately '0', so one variable has linear influence on calories. H0: B1 = 0 H1: B1 < 0 Portion has linear influence on calories.

I wanted to do Random Number Generation and Regression Analysis from the given data (which data is provided in our classroom) with Python but I could not find out our desired output for my programming lacking, so I try my best in excel to solve the problem.

Summary: We can find out our desired output by a statistical process for estimating the relationships among variables which is called Regression Analysis. By this analysis we know that portion has linear influence on calories. We also generate random numbers to generate a sequence that does not have any pattern, therefore appear to be random.