International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 10 ǁ October 2022 ǁ PP. 524-530

Application of Statistics in Solving Industry Based Challenges: A Case Study of Benedict Oil and Gas, Anambra State, Nigeria

[1]Jude Chukwura Obi and [2]Kosisochukwu Cynthia Okafor
[1, 2] Department of Statistics,
Chukwuemeka Odumegwu Ojukwu University, Anambra State, Nigeria
Corresponding Authors Email: obijudec@ymail.com

Abstract
The company, Benedict Oil and Gas has been studied, and data obtained thereof analyzed. The study aimed to underscore the usefulness of statistics in a service-oriented industry such as the Benedict Oil and Gas. Data analysis was tailored to find out if differences exist in the quantities of the major products (PMS, DPK and AGO) sold by the company. It sought to formulate a predictive model capable of predicting monthly sales, given different inputs. Lastly, it examined the possibility of associations among the products sold by the company. In the end, it was observed that there is marked differences in the quantities of the products sold by the company. The predictive model formulated was not found useful because all the assumptions underpinning the model were not met. Lastly, associations among the products sold were not useful because the correlation coefficients obtained were very low in most cases.
Keywords*: Benedict Oil and Gas, Analysis of variance, Predictive modelling, Regression, Correlation.*

Date of Submission: 08-10-2022 Date of acceptance: 20-10-2022

I. Introduction

Statistics is generally recognized for its usefulness in planning, research and innovation of tools for data analysis, thereby helping to meet with the organizational set goals. In view of this, it is overtly clear that statistics is useful in an industry such as the Benedict Oil and Gas, Anambra State, Nigeria. The Benedict Oil and Gas, as the name implies, majors in the sale of petroleum products, namely AGO (Automotive Gas Oil), DPK (Dual Purpose Kerosene) and PMS (Premium Motor Spirit). The core commitment of the company is sourcing the products usually from Nigerian National Petroleum Company (NNPC) and selling them to the final consumers. This business is involved and a challenging one too. Once supplies are received (which could be either AGO, DPK or PMS), somebody necessarily takes the reading to ensure that the volume of product ordered for, is actually what is supplied. Shortfalls, if found, are noted and proper record keeping must be maintained in order to follow-up business trends.

Record taking extends to daily volume of products sold, as well as the amount of money realized from sales of products. Record taking is a continuous exercise and the manager responsible for taking records diligently ensures that documentations are without errors, so as to avoid erroneous conclusions from data analysis. Some aspects of the record taking worthy of mentioning include the following:

- Volume of products sourced at departure point.
- Volume of products sourced at arrival point.
- Daily volume dispensed at each pump by a pump attendant.
- Daily cash sales accruing from the products sourced.
- Daily expenses incurred in managing the affairs of the company.

Information on volume of products sourced at both departure and arrival points will help the business owner understand what happens between both transiting points. It helps to answer the questions related to product theft while on transit.

The daily volume dispensed by each pump attendant will also help to understand if there is a correlation between the total volume supplied and the volume dispensed with by the pump attendants. For a healthy business, a perfect positive relationship is expected to exist, otherwise the observation made may be a sign or early warning that a remedial action is urgently called for. Total daily or monthly cash sales will provide additional information on the growth of the business. For a healthy business, the correlation between supplies and sales must be strongly positive. A perfect positive correlation is the best, but at least a strong positive correlation is acceptable.

Information on daily expenses incurred in the course of managing the business, is important to the management. If records of expenses incurred are consistently kept, over a period of time, a plot of what happens can help in early detection of trends. No business owner would delight in seeing surge in weekly or monthly expenses in the course of managing a business. A timely detection of surge signals a strong warning that a remedial action is urgently needed. It is important to note that all of theseinformation are possible if attention is paid to accurate record keeping, otherwise analysis of data will lead to misleading conclusions.

Thus far, industry-based application of statistics is heavily dependent on accurate and up to date record keeping. The staff whose duty is to keep the records must be given necessary training and where possible, a trained statistician should be engaged for this purpose.

II. Importance of the Study

This study is particularly important to Benedict Oil and Gas because it will help them achieve the following:

- Proper record keeping of quantities of product supplied, and daily volume sold.
- Ability to match product supply and daily volume of product sold, to be able to determine how much of a given supply is sold by a pump attendant. A good business should record at least 99% of total volume supplied as total volume sold by the pump attendant.
- Ability to engage a statistician to undertake the record keeping needs of the company.
- The ability to effectively utilize the result from data analysis in policy formulations in the company.

2.1 Aim and Objectives of the Study

The aim of the study is to underscore the usefulness of statistics in a service-oriented industry, such as the Benedict Oil and Gas. The various objectives include the following:

- To find out if there are differences in the quantities of PMS, DPK and AGO sold on monthly basis.
- To construct a regression model that is capable of predicting monthly sales given various inputs.
- To study possible association among the quantities of petroleum products sold by the company within the period under review.

III. Research Methodology

In this section, various statistical tools that shall be useful in the analysis of the datasets sourced from the Benedict Oil and Gas are going to be examined in considerable details. The tools will include the followings:

- Analysis of Variance
- Regression Analysis
- Correlation Analysis

It is hoped that by using these analysis tools, the aim and objectives of the study shall be met ultimately.

3.2 Analysis of Variance (ANOVA)

Analysis of variance(Scheffe, 1999) is a statistical procedure that is concerned with the comparison of means of more than two groups. It can be thought of as an extension of the t-test for two independent samples to more than two groups. Here, the purpose is to test for significant differences between class means, and this is accomplished by analyzing the variances.

When performing an ANOVA procedure, the following assumptions must be taken into consideration:

- The observations are independent of one another.
- The observations in each group come from a normal distribution.
- The population variances in each group are the same (homoscedasticity).

It is important to note that failure to comply with any of the afore-mentioned assumptions will lead to the use of nonparametric ANOVA test. A given ANOVA test may be one-way, two-way or factorial ANOVA. This study, however, will focus on the one-way ANOVA.

3.2.1 One-Way ANOVA

One-way analysis of variance is concerned with a procedure for testing more than two population means ($\mu_k \ s.t \ k \geq 3$) simultaneously(Obi, 2020). In this test procedure, interest is in the level of a single factor and we seek to know if the k-different groups involved in the study have the same mean or not. Importantly, the following hypotheses are tested:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$

H_1: At least two population means are not equal.

If the null hypothesis is rejected, it follows that at least a pair of means are not equal. The implication is that we carry out a post hoc test to find out the pair of means that gave rise to rejection of the null hypothesis.

A simple approach to analysis of variance is to obtain an F-statistic, which is a ratio of two variances given in (3.1).

$$F = \frac{\text{Between Sample Variance}}{\text{Within Sample Variance}} = \frac{ns_{\bar{x}_j}^2}{\mu_{s_j^2}};\ j = 1,2,\cdots,k. \quad (3.1)$$

The critical F for the test is given as follow:

$$F_{cri} = F_{\alpha,\ v_1,\ v_2}, \quad (3.2)$$

where $v_1 =$ the numerator degree of freedom $= k - 1$ and
$v_2 =$ Denominator degree of freedom $= \text{k(n} - 1) = \text{kn} - \text{k}$.
If $F < F_{cri}$, there is no sufficient reason to reject the null hypothesis. Otherwise, the null hypothesis is rejected.

Alternatively,the F-statistic (3.1) can be obtained by splitting the total variability into Sum of Squares Treatment (SSTr) and Sum of Squares Error (SSE). With that, one can obtain a ratio of two variances given as Mean Squares Treatment over Mean Square Residual. The equations that follow will explain further:

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{\bar{x}}\right)^2, \quad (3.3)$$

$$SSTr = n\sum_{i=1}^{k}(\bar{x}_i. - \bar{\bar{x}})^2, \quad (3.4)$$

$$SSE = SST - SSTr.$$

By making use of the various sum of squares, a corresponding F-statistic similar to (3.1) obtains as follows:

$$F = \frac{MSTr}{MSE} = \frac{SSTr/(k-1)}{SSE/k(n-1)}, \quad (3.5)$$

where

$MSTr =$ Mean Square Treatment

$MSE = Mean\ Square\ Error$

The critical Fvalue remains as stated in (3.2).The ANOVA table that gives further insight to (3.5) obtains as follows:

Table 1: Anova table for one-way analysis of variance

Source	SS	DF	MS	F
Treatment	SSTr	$k-1$	$\frac{SSTr}{k-1}$	$\frac{MSTr}{MSE}$
Error	SSE	$k(n-1)$	$\frac{SSE}{k(n-1)}$	
Total	SST	$kn-1$		

3.3 Regression Analysis

Regression analysis(Chatterjee & Hadi, 2006)concerns the use of a regression model which is solely for prediction purposes. It involves identifying the predictive relationship between a dependent variable and one or more independent variables. Here, we shall be concerned with more than one independent variable, hence, we focus on multiple regression. A multiple regression model is given in (3.6).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon. \quad (3.6)$$

Since we have up to three explanatory variables (independent variables) in this study, the number of p is 3, hence our model reduces to (3.7).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon. \quad (3.7)$$

Where:
Y is the output, dependent or response variable.

X,the independent, input, predictor or explanatory variable.
β_1, β_2, and β_3 are regression coefficients for variables X_1, X_2 and X_3 respectively.
$\beta_0$is the intercept point of the regression line, whereas$\epsilon$is the model's random error (residual) terms.

Estimate of the parameters of the model (3.7) is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{y} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}. \tag{3.8}$$

Hence, the model's estimate is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3. \tag{3.9}$$

It is important to note that the use of the regression model is contingent on satisfying some parametric assumptions. Otherwise, the predictions of the model will be unreliable.

3.4 Correlation Analysis

Correlation(Pearson, 1920), like covariance, is a statistical technique used to measure a relationship between two variables. The major weakness of covariance is that it lacks the ability to determine the strength of a relationship, but with correlation, the strength of a relationship can be determined. A population correlation coefficient is denoted with the symbol ρ, and lie between 0 and 1 inclusively. Hence, we write:

$$0 \leq \rho \leq 1.$$

When $\rho = 1$, we have a perfect positive correlation between the two variables involved. When it assumes the value 0, it means there is no relationship between the two variables in question. If the value of ρ is -1, we have perfect negative correlation between the two values in question. It should be noted that in practice, we do not usually use ρ in determining any association between any two variables, but alternatively we use a sample equivalence. In other words, ρ is estimated using sample correlation coefficient r. Similarly,

$$0 \leq r \leq 1.$$

Note that

$$r = \frac{cov(X_1, X_2)}{s_{x_1} s_{x_2}} \tag{3.10}$$

$$= \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}} \tag{3.11}$$

3.4.2 Hypothesis Testing

The null and alternative hypotheses are as follows:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

The test statistic is

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{3.12}$$

which has t-distribution with $n-2$ degree of freedom. The null hypothesis cannot be rejected if $|t| < t_{\frac{\alpha}{2},\ (n-2)}$.

IV. Data Presentation/Analysis

The data for analysis were collected on AGO (Table 4.1), DPK (Table 4.2) and PMS (Table 4.3). The data spans from January 2013 to December 2021. In other words, it is a monthly time series data.

Table 4.1: Quantities of AGO sold from January 2013 to December 2021

	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER
2013	11065.72	12453.72	12257.7	11652.4	13524.5	11435.7	10342.5	12343	13542.72	12800.09	13425.62	13425.74
2014	14325.81	17379.95	15848.8	16434.2	12275.2	13127.7	17123.2	15678.3	18947.08	15175.24	16425.24	18308.31
2015	15412.42	15475.75	16667.6	17505.7	16589.6	17417.2	15828.4	17227.8	18132.87	15882.66	18504.69	18383.55
2016	16418.26	18924.42	17745.7	16024.2	15978	14677	12706.6	15503.7	16377.57	17745.64	18802.14	18915.71
2017	17689.41	17678.91	18791.7	12141.3	18148.9	15081.1	17689	18111.5	15347.44	16255.35	17759.61	19503.41
2018	18414.43	17667.57	11414.4	16243	18332.3	19774.9	12339.6	13013	16667.38	18552.79	17553.73	18849.61
2019	13146.86	17683.98	17187.2	16303.8	11374.3	13390.1	14219.3	15900.4	12356.01	15935.54	18308.29	20849.64
2020	19426.56	19751.81	18164.9	17342.6	19061.4	18535.1	17876.9	16331.8	15403.81	18384.09	19506.31	20316.4
2021	20540.76	20343.26	19917.1	20247	21427.8	21817.3	20953.5	21456.7	21838.01	20211.04	22731.37	23362.32

Table 4.2: Quantities of DPK sold from January 2013 to December 2021

	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER
2013	3019.99	3289.86	3738.61	3813.2	3298.2	3159.8	3264.4	2899.19	3013.74	3189.37	3500.12	2998.78
2014	4038.25	4296.62	4358.79	4160.7	4511.5	4558.1	4632.1	4084.04	4285.36	4310.01	4810.03	4551.18
2015	4325.09	4532.75	4678.83	4332.2	4596.2	4763.2	4865.2	4121.71	4899.07	4724.11	4843.78	4910.65
2016	3951.31	3865.13	3716.47	3637.3	3571.3	3722	3417.1	3284.75	3737.43	3995.8	3843.03	3991.21
2017	3538.47	3422.4	3639.16	3880	3919.8	3401.6	3312	3209.35	3308.5	3711.93	3652.77	3466.81
2018	3389.78	3422.05	3510.11	3613.5	3430.8	3338.8	3278.7	3583.69	3633.86	3152.98	3259.11	3071.72
2019	3389.24	3459.21	3589.31	3643.5	3689.2	3551.8	3661.7	3425.27	3351.23	3281.32	3159.44	3539.12
2020	4628.49	4529.66	4651.17	4828.3	4739	4947.1	4581.8	3589.46	3283.13	3331.34	4450.44	3256.73
2021	3681.54	3981.28	3954.88	3954.4	3781.8	3829	3821.1	3957.24	3173.53	3191.52	3106.96	3783.13

Table 4.3: Quantities of PMS sold from January 2013 to December 2021

	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER
2013	134051.7	122237.58	122435.9	123521.1	124051.5	124452.2	125521.2	125305.8	125409.81	125550.3	127899.38	128501.98
2014	129051.3	129802.78	128149.8	128536.4	130098.4	131059.8	130210.8	130370.3	131607.8	131449.1	131502.96	131789.69
2015	131890	130095.56	139657.3	138390	137213.8	137114.2	136756.8	139788.4	137567.97	139895.4	138303.19	139809.79
2016	172993.6	140159.11	140261.9	140269.7	141044.4	140954.9	142256	140318.8	141196.59	141398.3	141596.71	141678.14
2017	143789.8	145439.96	144814.2	145362.2	142158.8	143998.2	144598.6	149132.7	145321.31	143067	142969.84	146389.51
2018	173898.2	147811.27	178100.7	147391.3	148421	146455.3	147418.5	149934.2	145695.13	148943.1	179280.17	179978.05
2019	149321.7	148963.51	149993.2	149837.8	149989.6	150259.8	150079.5	150001.8	150098.79	151028.8	151029.21	152346.81
2020	150039.9	150939.05	150179.1	152013.7	151013.6	150007.9	145980.6	155932.1	152397.35	153397.5	152503.95	151731.63
2021	154022.5	155193.62	153759.6	156755.2	155671.2	157666.2	156523.1	155779.8	156885.24	157645.8	158116.45	158981.18

Table 4.4: Amount in Naira realized from sale of products from January 2013 to December 2021.

	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOMBER	NOVEMBER	DECEMBER
2013	18883291	17791384.2	17881152	17913529	18214421	17833798	17778436	18048675	18315454	18231612	18703821	18660847.6
2014	19248873	19996612.7	19514589.5	19627604	19088234	19381012	20076021	19681881	20516103	19767471	20137684	20477166.5
2015	23581125	23396210	25070816.5	24954895	24675705	24881038	24534122	25037529	25134315	24958017	25300558	25514242.3
2016	30045662	25835226.6	25534429.2	25115867	25197769	24930752	24574750	24897513	25361576	25782997	26008924	26091306.6
2017	26729426	26931668.5	27179818.8	25746718	26727959	26087753	26772047	27509569	26315829	26337249	26663785	27525181.3
2018	32778145	28559577.4	31722834.4	28207263	28823408	28846259	27115074	27783171	28057780	28859451	33350138	33714819.5
2019	27880754	29074742.2	29145739.3	28902109	27610676	28148729	28383135	28742240	27774366	28860523	29458570	30481850.4
2020	32204525	32415867.3	31856943.3	31974966	32294310	32043322	31053492	31884646	30915801	31991790	32572699	32270494.5
2021	34567785	34823813.2	34419768.6	35034237	35208709	35701300	35193834	35316124	35285027	34824764	35784296	36461550

The analysis of data in this section will be tailored to provide answers to the research objectives. For this reason, the research objectives will be revisited one after the other.

4.1 Differences in the Quantities of AGO, DPK and PMSSold on Monthly Basis

In order to find out if there are differences in the quantities of AGO, DPK and PMS sold on monthly basis, analysis of variance test will be carried out on the datasets of Tables 4.1 to 4.3. The null and alternative hypotheses are:

$$H_0: \mu_{AGO} = \mu_{DPK} = \mu_{PMS}$$
$$H_1: \text{At least a pair of means are not equal.}$$

To ensure that there is compliance with the assumptions of ANOVA, a Shapiro normality test carried out in R(SHAPIRO & WILK, 1965), shows that normality assumption is rejected for DPK (p-value = 0.00004383) and PMS (p-value = 0.0007454). In the case of AGO, assumption of normality could not be rejected (p-value = 0.08124). Since two datasets failed to comply with the normality assumption, a nonparametric Kruskal-Wallistest('Kruskal-Wallis Test', 2008)shall be used. The test shows that at a p-value less than 2.2e-16, the null hypothesis is rejected. The rejection of the null hypothesis requires that a post-hoc Dunn's test(Dunn, 1961) is carried out. The outcome of the test in R is contained in Table 4.5.

Table 4.5: Output of post-hoc Dunn's test in R

	Comparison	Z	P.unadj	P.adj
1	AGO - DPK	8.472217	2.407660e-17	2.407660e-17
2	AGO - PMS	-8.472217	2.407660e-17	3.611490e-17
3	DPK - PMS	-16.944434	2.115601e-64	6.346803e-6

Note that Table 4.5 shows that for each paired comparison, both p-value unadjusted and adjusted are all less than 0.05. In that case, the null hypothesis stands totally rejected.

4.2 Prediction of Monthly Sales

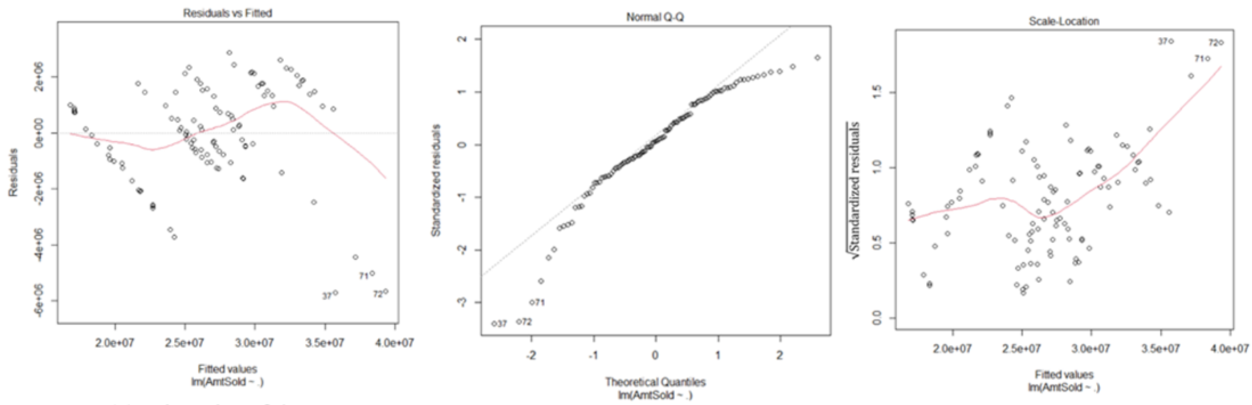
The model for prediction of monthly sales is given in (4.1)

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \epsilon \quad (4.1)$$

Using the R statistical software, estimate of the model is

$$\hat{Y} = -30{,}040{,}000 + 623.4X_1 + 149.2X_2 + 317.7X_3. \quad (4.2)$$

On compliance with the model's assumptions, Figure 4.1 (a) is the residual vs fitted plot. It helps to show if the linearity assumption of data is met. So far, a pattern can be observed in the plot which is a sign that linearity assumption is violated. Regarding the normal q-q plot (Figure 4.1 (b)), normality assumption of residuals is violated as well because data points are not spreading continuously along the diagonal line. Lastly, on assumption of homogeneity of variance, the scale location plot (Figure 4.1 (c)) shows that the red line is not at least approximately horizontal. Again, data points scatter randomly which indicates violation of homogeneity assumption. It is important to note that transformation of the observed values (Y) could not improve compliance with the model's assumptions. Hence, the regression model being proposed should be discarded because no reliable prediction could obtain, particularly as the model's assumptions have all been violated.



(a) Linearity of data (b) Normality of residual plot (c) Homogeneity of variance plot

Figure 4.1: Various plots for verification of compliance with the model's assumptions.

4.3 Association Among Petroleum Products Sold

To find out if there are associations among petroleum products sold by the company within the period under review, a correlation analysis was carried out. Using the cor_mat() function in R, from the library rstatix(Kassambara, 2021), Table 4.6 that follows was obtained.

Table 4.6: Correlation coefficient among petroleum products sold by the company.

Rowname	AGO	DPK	PMS
<chr>	*<dbl>*	*<dbl>*	*<dbl>*
AGO	1	0.19	0.48
DPK	0.19	1	-0.2
PMS	0.48	-0.2	1

Based on Table 4.6, there is weak positive correlation between AGO and DPK, mild positive correlation between AGO and PMS and lastly, weak negative correlation between DPK and PMS. Thus, it is not possible that the quantity sold of any given product can be used as a bases of understanding how another product sales.

V. Summary/Conclusions

Based on analysis of data carried out in section 4, the quantities of AGO, DPK and PMS sold by the company on monthly basis sharply differ from one another. The highest sold is PMS (15569633.7), followed by AGO (1806789.3) and lastly DPK (412028.7). The company should invest more in PMS, followed by AGO because the two products sell faster and it translates to improved revenue to the company.

On the prediction of sales given different quantities of products sold, the model estimate given in (4.2) has a major drawback. It is the inability to comply with all the assumptions underpinning the use of the model. For this reason, the model should be discarded and effort intensified to examine the procedure for recording data by the company. If data are correctly recorded, it is possible that a predictive model that conforms with all required assumptions can be formulated.

Regarding association among the quantities of different products sold, a correlation analysis was carried out. It shows that there is weak positive correlation between AGO and DPK $(r = 0.19)$, mild positive correlation between AGO and PMS $(r = 0.48)$ and weak negative correlation between PMS and DPK (-0.2). On the strength of the correlation coefficients, it is not possible to use product to understand what goes on with another product in terms of the quantity sold.

References

[1]. Chatterjee, S., & Hadi, A. S. (2006). Regression analysis by example. John Wiley & Sons.

[2]. Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56(293), 52–64.

[3]. Kassambara, A. (2021). rstatix: Pipe-Friendly Framework for Basic Statistical Tests. https://CRAN.R-project.org/package=rstatix

[4]. Kruskal-Wallis Test. (2008). In The Concise Encyclopedia of Statistics (pp. 288–290). Springer New York. https://doi.org/10.1007/978-0-387-32833-1_216

[5]. Obi, J. C. (2020). A Foundation Course in Statistics with Applications in R. Favour Fountain Concepts.

[6]. Pearson, K. (1920). Notes on the history of correlation. Biometrika, 13(1), 25–45.

[7]. Scheffe, H. (1999). The analysis of variance (Vol. 72). John Wiley & Sons.

[8]. SHAPIRO, S. S., & WILK, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591