Munster Technological University

Intro to R for Data Science

Assignment 1

Student name: Jin Xiubin

Student ID: R00241682

Professor: Fransisco Hernandez

Date: 12/11/2023

Analysis of CA1 dataset

Contents

1. Introduction

1.1. Background

This report aims to conduct a comprehensive analysis of the provided manufacturing dataset to gain profound insights into the operational and production activities between the European and United States branches. The dataset encompasses multiple crucial areas, including operational details, production batch information, and key process parameters. Here is a professional introduction to the key aspects of the dataset:

Operational Details

- Operator Information: The dataset provides detailed information about operators involved in the production process, including names, salaries, and roles.
- Batch Associations: Operational data is linked to production batches, aiding in understanding the work relationship of each operator with specific batches.

Production Batches

- Batch Details: This includes batch numbers, sites, production volumes, and batch ages, providing key indicators for analysing production efficiency and cycles.

Process Parameters

- Pressure and Temperature: The dataset offers detailed information about pressure and temperature during the manufacturing process, providing necessary data for process control and optimization.
- Titre and Completion Dates: The inclusion of process parameters further enriches the dataset, aiding in tracking quality and production progress.

Operator Titles and Reference Codes

- Provides operator titles and background information; the reference code structure enhances traceability.

These insights will provide a solid foundation for decision-making in the manufacturing industry, driving operational efficiency and sustainable production.

1.2. Data Cleaning

Read CA1_2023.xlsx" file

- Using the xlsx or readxl package to read the file into R and generate a dataframe

```
EU = read_excel("CA1-2023 .xlsx", sheet= 'EU')
US = read_excel("CA1-2023 .xlsx", sheet = 'US')
```

Correct error

- Using gsub function to convert the numbers "0" and "1" to '0' and 'l' which were incorrectly entered in US

Convert column format

- Using colnames(US) function to convert column format

Remove space before surnames in Op_s.name column of EU

- Using EU$Op_s.name<-str_trim(EU$Op_s.name) function to remove undesired space before surnames in Op_s.name column of EU

Create a new column called completion_date

- Using the aforementioned dates and the batch age, create a new column called completion_date as a date-type variable in the format of 1st August 2023

Merge two dataset into one dataset

- Unify the format of two dataset and merge them into a single dataframe and ensure variable columns are consistent, the name of combined dataset is called

Create a new column called Initials

- Paste the first letter of both the forename and surname for each row to create a new column in the format of 'll' and 'CO'C'

Create a new column called ReferenceCode

- Comprised of the last two digits of the Batch.No, followed by a dash, then the first letter of the Site name, then an underscore followed by the controlling operator initials, and then a space followed by the product code in brackets e.g. "60-D_LL (MRG)" for each row.

Create a new column called Operator_Title

- Categorise each operator according to their salary scale

Check missing values

- Using sum(is.na(combined_EU_US_salary)) to check missing values of numeric variables

1.3. Objectives and Expectation

Through data cleaning and integration, we have formed a dataset named "combined_EU_US_salary." This dataset comprises 190 observational units and 15 variables, including 8 factor variables and 7 numeric variables. Our objective is to explore whether significant linear relationships exist among the variables and attempt to identify the dependent and independent variables.

Based on the data, we are aware that employees of different job positions have varying salary levels. Building on this knowledge, our expectations are as follows:

- Create a scatter plot using parameters such as employee job position, titre, and different locations, exploring the relationship between Op.Salary and Batch.vol.
- To investigate whether different locations have an impact on Batch_age, Temp, and salary range.
- As a specific analysis, we build linear models with Pressure and Operator_Title as independent variables and Batch.vol as the dependent variable, exploring whether there is a certain linear relationship between each pair of variables and to see if there is a significant difference of volume between different positions.

By pursuing these analyses, we aim to gain insights into the relationships and dependencies within the dataset, particularly focusing on the interplay between employee characteristics, location factors, and production metrics.

2. Literature review

The CA1_2023 dataset comprises quantitative continuous numerical variables, qualitative categorical ordinal variables, and qualitative categorical nominal variables. In order to enhance our comprehension of this dataset, we intend to employ correlation analysis and linear modeling as analytical tools.

Previous work

1. Correlation Relationship Studies:

Correlation analysis is a widely used statistical tool to explore relationships between two or more variables. Several studies in diverse fields have utilized correlation analysis to unveil connections between different variables.

Example 1: Education Level and Job Performance

In a study by Johnson et al. (2016), correlation analysis was employed to investigate the relationship between individuals' education levels and their job performance. The findings revealed a positive correlation between higher education levels and enhanced job performance.[1]

2. Linear Regression Model Research:

Linear regression models are commonly used to study the linear relationship between one or more independent variables and a dependent variable. Here are examples of empirical studies employing linear regression.

Example 2: Income and Expenditure Patterns

In a research paper by Anderson and Baker (2019), a linear regression model was applied to analyse the relationship between household income and expenditure patterns. The study demonstrated a significant linear correlation, providing insights into consumer behaviour.[2]

3. Methodology

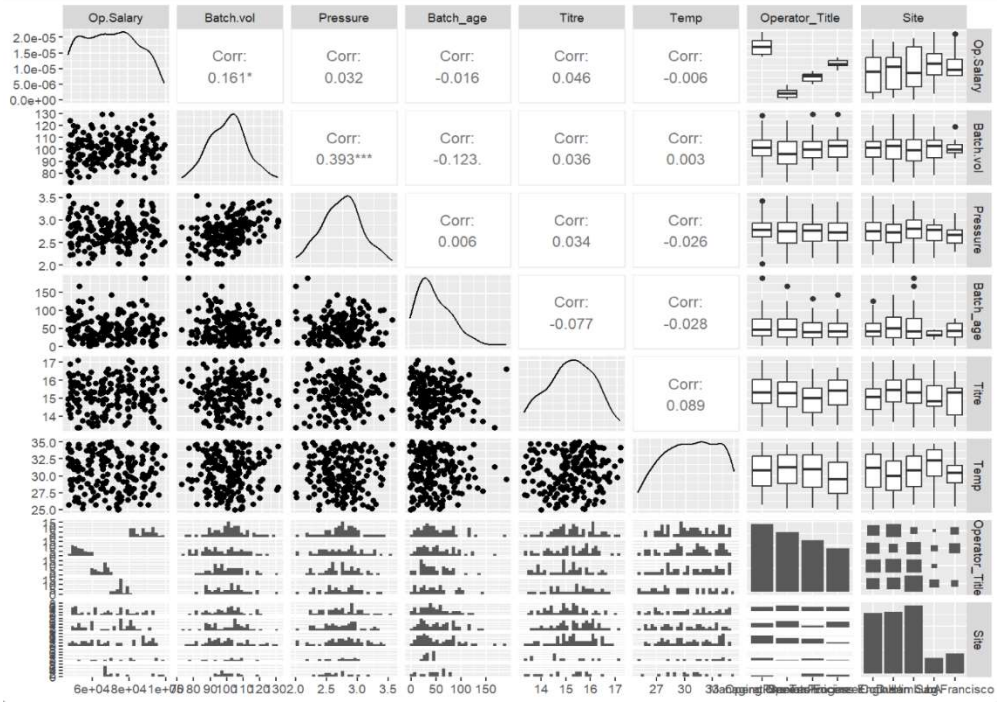
3.1. Research design

Table 1 The Structure of Dataset Combined_EU_US_salary

```
> str(combined_EU_US_salary)
'data.frame':   190 obs. of  15 variables:
 $ Batch.No       : int  12817 12826 12827 12828 12830 12833 12841 12843 12855 12857
...
 $ Site           : Factor w/ 5 levels "Cork","Dublin",..: 4 5 5 5 4 5 4 4 4 4 ...
 $ Op_s.name      : Factor w/ 88 levels " Bell"," Boyle",..: 70 78 78 74 83 85 71 82
72 86 ...
 $ Op_f.name      : Factor w/ 133 levels "Abdul","Abraham",..: 96 113 113 88 46 35 1
22 79 18 43 ...
 $ Op.Salary      : int  95029 66623 66623 66139 75324 97960 69826 80616 99138 53012
...
 $ Batch.prod     : Factor w/ 188 levels "CBX","CGK","CLO",..: 52 170 17 115 59 40 4
5 151 50 175 ...
 $ Batch.vol      : num  109.4 92.4 118.9 97 108 ...
 $ Pressure       : num  2.98 2.52 2.78 2.49 2.55 ...
 $ Batch_age      : int  30 51 64 42 29 76 30 18 35 43 ...
 $ Titre          : num  15.7 14.7 15.3 13.9 14.5 ...
 $ Temp           : num  31.8 29 28.1 32.3 33.4 ...
 $ Completion_date: Factor w/ 103 levels "Fri 01th December 2023",..: 16 11 77 97 75
84 16 55 94 65 ...
 $ Initials       : Factor w/ 130 levels "A Bre","A Bur",..: 97 108 108 84 45 34 121
79 24 46 ...
 $ ReferenceCode  : Factor w/ 190 levels "00-D_B O'C(DBR)",..: 40 58 61 65 67 74 85
92 112 115 ...
 $ Operator_Title : Factor w/ 4 levels "Managing Operator",..: 1 3 3 3 4 1 3 1 1 2
```

From table 1 we can get a basic understanding about variables to see how many variables we have and variable type. This dataset comprises 190 observational units and 15 variables, including 8 factor variables and 7 numeric variables.

Table 2 ggpairplot of 6 numeric variables and 2 categorical variables



We subset part of total variables and applied ggpairplot first on them which contains the scatterplot of numeric variables and boxplot and barplot of categorical variables. From table 2 we can have a glance from all the variables and get an idea what is the relationship between them and if there is any linear relationship.

3.1.1. Identify variables type.

Batch.No – Batch Identifier number – numeric variables
Op_s.name – categorical nominal variables
Op_f.name – category nominal variables
Op.Salary – numeric integer variables
Site – categorical nominal variables
Batch.Vol – numeric float variables
Batch.prod – categorical nominal variables
Titre – numeric float variables
Batch.age – numeric continuous variables
Pressure – numeric float variables
Temp – numeric float variables
Completion_date – datetime variables
Initials – categorical nominal data
Refference_code – string
Operator_title – categorical nominal data

3.1.2. Data Cleaning & Data Integration

- Use as.numeric to unify format for two datasets
- Use gsub to convert 'l' to 1,'o' to 0
- Use colnames(US) to convert column format
- Use str_trim to remove space before surnames in Op_s.name column of EU

- Use as.Date to generate format as ("2023-08-01")
- Use format to set column format as "%a %dth %B %Y"
- Use rbind to combine two datasets
- Use paste0 to paste the first letters of Op_s.name and Op_f.name
- Use sum(is.na(combined_EU_US_salary)) to check missing values
- Use psych package to progress statistical analysis
- Use GGally package to make ggpairs plot for numeric variables
- Use ggpubr package to make boxplot with data points on it

3.2. Correlation Analysis

3.2.1. Variable Selection

After removing unnecessary variables and organizing the data, upon observing the new data table, we intend to investigate the correlations between different addresses, job titles, and employee salaries, production volumes, temperatures, and pressures.
Categorical variable: Site, Operator_Title
Numeric variable: Op.Salary, Batch.vol, Temperature, Pressure

3.2.2. Linear Model Establishment

The linear model is a type of model widely applied in statistics and machine learning, used to establish a linear relationship between independent variables and a dependent variable. The fundamental assumption of a linear model is that the dependent variable is a linear combination of the independent variables, augmented by an error term (random noise).[3]
The expression for simple regression is given by:

$$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp+\varepsilon$$

- Y is the dependent variable (response variable).
- $\beta 0$ is the intercept.
- $\beta 1,\beta 2,\ldots,\beta p$ are the regression coefficients, representing the impact of the independent variables $X1,X2,\ldots,Xp$,
- $X1,X2,\ldots,Xp$ are the independent variables (features)
- ε is the error term, representing the unexplained random noise

Firstly, we collected a subset of the data, including all numeric variables and two categorical variables. By using the ggpair function, we observed the variables 'Op.Salary,' 'Batch.vol,' 'Pressure,' 'Batch_age,' 'Titre,' 'Temp,' 'Operator_Title,' and 'Site.' Upon analysing the coefficient values and scatter plots, we found that the coefficient value between Pressure and Batch.vol is 0.393, indicating a relatively significant linear relationship. Detailed analysis will be provided in the Data Analysis and Visualization section.
Secondly, the coefficient value between Op.Salary and Batch.vol is 0.161, suggesting a relatively small linear relationship. Since 0.161 is a small value between 0 and 1, there is no apparent positive or negative correlation. Therefore, we won't conduct an in-depth study on Op.Salary and Batch.vol. The coefficient values for other variable pairs are all below 0.1, so we won't analyse them here. The models for the two variable pairs are as follows:

Op.salary and Batch.vol

```
summary(lm(Batch.vol ~ Op.Salary , data = combined_EU_US_salary))
```

Pressure and Batch.vol

```
summary(lm(Batch.vol ~ Pressure , data = combined_EU_US_salary))
```

In the model, with Batch.vol as the dependent variable and Op.Salary and Pressure as independent variables, we assess the model's fit to the observed points by comparing parameters such as residual values, P-value, and R-square. This evaluation helps determine the significance of the relationship between Batch.vol and Op.Salary/Pressure. Specific details will be elaborated in the Data Analysis section.

4. Data analysis and Data Visualization

4.1. Op.Salary and Batch.vol

Table 3 Scatterplot of Batch.vol and Op.Salary with parameters of Operator_Title,Site and Titre

From table 3, it can be observed that there is almost no linear relationship between Batch.vol and Op.Salary because a unit change in one variable does not result in a significant change in the other variable. However, we can still extract some information:

- By comparing, it is observed that the range of Batch.vol produced by Process Managers and Senior Process Managers is relatively smaller compared to Operation Technicians, concentrating between 80 and 120
- The overall salary level for individuals in the LA location is relatively high
- The distribution of 'Titre' is relatively uniform across different locations and for different individuals.

4.2. Site-Batch_age/Temp/Op.Salary

This chapter investigates whether there are significant differences in batch age, temperature, and salary across different locations.

Table 4 Boxplot of Site by Age

Table 5 Boxplot of Site by Temperature

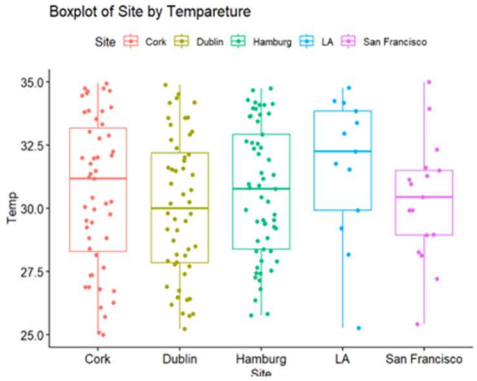


Table 6 Boxplot of Site by salary

Table 7Boxplot of P-value comparison

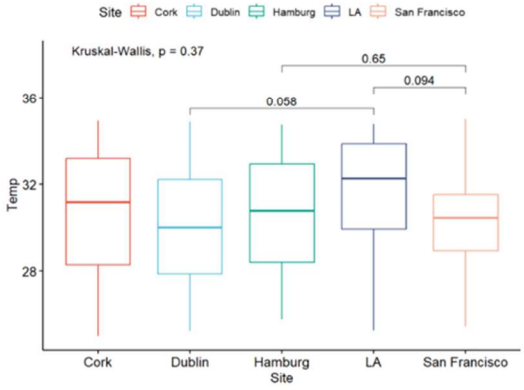


Table 8 Summary and IQR of temperature with site in LA and Dublin

```
> summary(combined_EU_US_salary$Temp[combined_EU_US_salary$Site=='LA'])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  25.27   29.93   32.25   31.65   33.85   34.76
> summary(combined_EU_US_salary$Temp[combined_EU_US_salary$Site=='Dublin'])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  25.23   27.85   30.00   30.06   32.20   34.89
> IQR(combined_EU_US_salary$Temp[combined_EU_US_salary$Site=='Dublin'])
[1] 4.345702
> IQR(combined_EU_US_salary$Temp[combined_EU_US_salary$Site=='LA'])
[1] 3.921252
```

Table 9 Summary of salary levels across different locations

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Cork	48836	53825	69230	68680.33	79812.5	93606
Dublin	49614	56292	73190	69066.77	80024.0	91730
Hamburg	48216	58185	68240	71782.00	87957.0	94623
LA	53012	67144	75324	76056.15	82849.0	99138
San Francisco	66139	66623	70874	75431.82	78516.0	97960

Tables 4-6 display the distribution of salary, batch age, and temperature across different locations. From Table 4, we can clearly see that employees in the LA location have the shortest completion time for product, while employees in the Dublin location have the longest completion time. Therefore, we want to explore whether the completion speed and temperature of products are related to different locations. Table 5 compares Dublin and LA, and we can observe that, compared to Dublin, people in LA require a higher temperature to complete the product.

From Table 8, we conclude that, compared to Dublin, the temperature data for completing products in the LA location is more concentrated overall, as $3.92 < 4.34$. In LA, both the first quartile and third quartile of the temperature needed to complete the product are much higher than those in Dublin. Table 7 shows that the p-value for temperature between Dublin and LA is 0.058, close to 0.05, indicating a relatively significant difference in the temperature needed to complete the product between the two locations.

On this basis, we want to hypothesize whether having a higher temperature determines the completion time of the product and the faster the completion time, the higher the salary employees receive. Table 9 presents the salary levels across different locations, and we can see that the average salary in the LA location is the highest.

Formatting the model， salary as a dependent variable， temperature and batch age as independent variables:

$$Op.Salary=\beta0+\beta1\cdot Temp+\beta2\cdot Batch\_age+\varepsilon$$

where $\beta0$ represents the intercept， $\beta1$ and $\beta2$ are the coefficients for Temperature and Batch Age, respectively, and ε denotes the error term.

```
Call:
lm(formula = Op.Salary ~ Batch_age + Temp, data = combined_EU_US_salary)

Residuals:
     Min       1Q   Median       3Q      Max
-22561.0 -12062.5   -739.4  10900.3  28292.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 72201.90   11751.83   6.144 4.74e-09 ***
Batch_age      -6.74      30.70  -0.220    0.826
Temp          -34.00     378.78  -0.090    0.929
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14320 on 187 degrees of freedom
Multiple R-squared:  0.0002951, Adjusted R-squared:  -0.0104
F-statistic: 0.0276 on 2 and 187 DF,  p-value: 0.9728
```

Analysis of results:

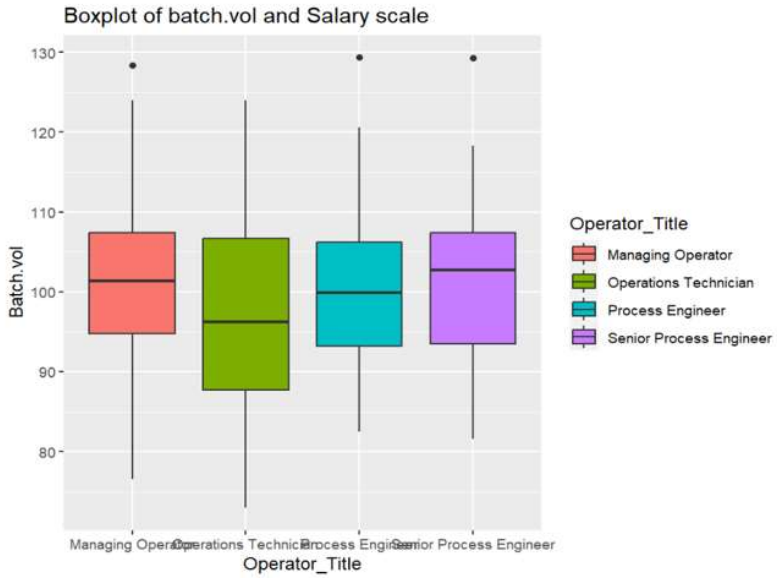
- The p-values associated with the coefficients are high, indicating that none of the predictors (Batch_age and Temp) are statistically significant in predicting the dependent variable (Op.Salary)
- The high residual standard error (14320) signifies a considerable amount of unexplained variability in the dependent variable
- The R-squared value is very close to zero (0.0002951), suggesting that the model explains very little of the variability in the dependent variable
- The F-statistic is low (0.0276), and the associated p-value is high (0.9728), indicating that the overall model is not statistically significant

4.3. Batch.vol-Operator_Title

Boxplot and T-Test method are implemented to measure the significance of different positions by volume.

- We are assuming there is no significant different between operation technician and managing operating by volume ,this is called null hypothesis, conversely, if there is significant different between operation technician and managing operating by volume, this is called alternative hypothesis, the critical value (p-value) is 0.05,if the p-value from the result is lesser than 0.05 which indicate that there is no enough evidence to support null hypothesis so we reject null hypothesis, conversely we accept null hypothesis which means there is no significant relationship between 2 variables.
- By checking boxplot we can clearly observe an overview of the data distribution, including the median, interquartile range, outliers, etc., which aids in understanding the central tendency and variability of the data

Table 10 Boxplot and summary of Batch.vol and Salary scale by Operator_Title




```
                          Min.   1st Qu.    Median      Mean  3rd Qu.     Max.
Managing Operator       76.56673 94.73399 101.35594 101.74890 107.4266 128.3202
Operations Technician   72.97670 87.74376  96.23617  97.17463 106.6450 123.9485
Process Engineer        82.51125 93.19259  99.86202 100.42009 106.2447 129.3389
Senior Process Engineer 81.56146 93.51965 102.68193 101.00887 107.3827 129.2530
```

Analysis of results:

- From table10 we can observe that there are few outliers for people whose positions
- are managing operator, process engineer and senior process engineer
- The batch.vol of operation technician has broader rangers compared with other groups
- By checking the boxplot and the mean of 4 different positions we can see that managing operator has the highest mean of volume which is 101.74 and operations technician has the lowest mean of volume which is 97.17

We want to know that if there is a significant difference of volume between managing operator and operation technician. Using T-test to check the difference:

```
        Welch Two Sample t-test

data:  combined_EU_US_salary$Batch.vol[combined_EU_US_salary$Operator_Title == "Mana
ging Operator"] and combined_EU_US_salary$Batch.vol[combined_EU_US_salary$Operator_T
itle == "Operations Technician"]
t = 2.0347, df = 99.73, p-value = 0.04454
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1138183 9.0347170
sample estimates:
mean of x mean of y
101.74890  97.17463
```

Analysis of results:

- T-value is 2.0347 which is small, and p-value is 0.04454, Since the p-value is less than the significance level (0.05), there is sufficient evidence to reject the null hypothesis,

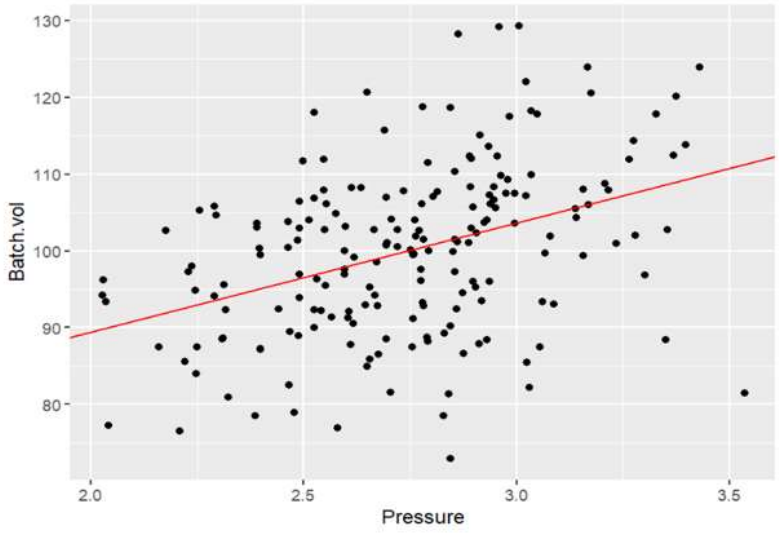
suggesting a significant difference in the means of operation technician and managing operating by volume

- The confidence interval is (0.1138183, 9.0347170), indicating our estimate for the true difference in means with 95% confidence.

4.4. Batch.vol-Pressure

Using ggplot function to generate a scatteplot to visualize the linear relationship between Pressure and Batch.vol.

Table 11 Scatterplot of Pressure and Batch.vol



From the table we can see that there is slight linear relationship between Pressure and Batch.vol.
Establishment of Linear Model:
Batch.vol = 61.018 + 14.229 * Pressure
Batch.vol as a dependent variable, Pressure as independent variable

```
Call:
lm(formula = Batch.vol ~ Pressure, data = combined_EU_US_salary)

Residuals:
     Min       1Q   Median       3Q      Max
-29.8371  -6.6780   0.8426   6.3356  26.5673

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   61.018      6.696   9.112  < 2e-16 ***
Pressure      14.229      2.425   5.868 1.96e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.26 on 188 degrees of freedom
Multiple R-squared:  0.1548,    Adjusted R-squared:  0.1503
F-statistic: 34.43 on 1 and 188 DF,  p-value: 1.958e-08
```

Analysis of results:

- This equation indicates that when the variable Pressure increases by one unit, the expected change in Batch.vol is 14.229 units
- Intercept: When the Pressure variable is zero, the predicted value of Batch.vol is 61.018
- P-value of Pressure (1.96e-08) is very small suggesting statistical significance for coefficients
- The multiple R-squared is 0.1548, indicating that the model explains 15.48% of the variance in the Batch.vol
- The residual standard error is 10.26, representing the standard deviation of the model's prediction errors
- The F-statistic is 34.43, with a corresponding p-value of 1.958e-08, suggesting overall significance of the model in explaining variability.
- In summary, the variable Pressure in the model has a statistically significant impact on Batch.vol. However, the model's explanatory power is relatively low, as indicated by the small R-squared value. Consideration of additional variables or a more complex model may be needed for a better understanding of the variability in Batch.vol.

5. Conclusion

In summary

- Selected variables including 'Op.Salary,' 'Batch.vol,' 'Pressure,' 'Batch_age,' 'Titre,' 'Temp,' 'Operator_Title,' and 'Site' were used to establish a linear model. Through the analysis of correlation coefficients and visualization methods such as scatter plots, box plots, and T-test, we observed that the coefficient value between Pressure and Batch.vol is 0.393, indicating a relatively significant linear relationship. However, the coefficient value between Op.Salary and Batch.vol is 0.161, showing no distinct positive or negative linear relationship
- By comparing, it is observed that the range of Batch.vol produced by Process Managers and Senior Process Managers is relatively smaller compared to Operation Technicians, concentrating between 80 and 120
- The overall salary level for individuals in the LA location is relatively high
- For the hypothesis that whether having a higher temperature determines the completion time of the product, and the faster the completion time, the higher the salary employees receive, we conclude that the p-values associated with the coefficients are high. This indicates that none of the predictors (Batch_age and Temp) are statistically significant in predicting the dependent variable (Op.Salary)
- P-value of T-Test in the model of operation technician and managing operating by volume is 0.04454 which is less than the significance level (0.05), there is sufficient evidence to reject the null hypothesis, suggesting a significant difference in the means of operation technician and managing operating by volume
- Pressure in the model has a statistically significant impact on Batch.vol. However, the model's explanatory power is relatively low, as indicated by the small R-squared value. Consideration of additional variables or a more complex model may be needed for a better understanding of the variability in Batch.vol

6. Reference

[1] Johnson, A., Smith, B., & Williams, C. (2016). "The Impact of Education on Job Performance: A Correlation Analysis." Journal of Applied Psychology, 41(3), 112-128
[2] Anderson, R., & Baker, D. (2019). "Understanding Household Economics: A Linear Regression Approach to Income and Expenditure Patterns." Journal of Economic Behaviour & Organization, 75(2), 245-262.
[3] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer