**Data Science**

**Home Work 1.**

**RE6124019 吳明軒**

1. State the Data Science Work of Jeff Hammerbacher’s Model.
2. Identify Problem: This is where you define the problem you want to solve or the question you want to answer using data. It involves understanding the business or research context and formulating a clear problem statement.
3. Instrument (Select and Organize) Data Sources: In this step, you identify the data sources that are relevant to your problem. You might need to decide what data to collect, whether it's from databases, APIs, sensors, or other sources. Organizing data sources includes setting up data pipelines or connections to retrieve data.
4. Collect Data: This is the process of actually gathering the data you identified in the previous step. Data collection can be a complex and time-consuming process, involving data extraction, scraping, or querying databases.
5. Prepare Data (Integrate, Transform, Clean, Filter, Aggregate): Once you have the data, you need to make it ready for analysis. This involves various tasks such as data cleaning (removing duplicates, handling missing values), data transformation (e.g., converting data types), data filtering (removing irrelevant data), and data aggregation (if needed for your analysis).
6. Build Model: In this step, you create the machine learning or statistical model that will help you address the problem. Depending on the nature of the problem, this could involve various algorithms and techniques for classification, regression, clustering, etc.
7. Evaluate Model: After building the model, you need to assess its performance. This involves using evaluation metrics and techniques to determine how well the model is performing. You may need to iterate on the model and make improvements based on the evaluation results.
8. Communicate Results: Finally, you need to communicate your findings and results to stakeholders. This often includes creating reports, visualizations, or presentations that convey the insights gained from the data analysis and the implications for decision-making.
9. State the three basic principles in design of experiment. Explain how and why to use them.
10. Randomization:

* How to use it: Randomization involves randomly assigning subjects or experimental units to different treatment groups. This helps eliminate bias and ensures that each group is representative of the overall population.
* Why to use it: By randomly assigning subjects, you reduce the likelihood of systematic bias or confounding variables affecting the results. This ensures that the results can be generalized to a larger population.

1. Replication:

* How to use it: Replication involves conducting the same experiment multiple times under similar conditions. Each replication is an independent repetition of the experiment.
* Why to use it: Replication helps verify the consistency of results. If similar results are obtained in multiple replications, it increases confidence in the findings and reduces the impact of random variation.

1. Control:

* How to use it: Control involves creating a control group that does not receive the treatment or intervention being tested. This group is used as a baseline for comparison with the experimental group(s).
* Why to use it: Control groups help establish causality by allowing researchers to isolate the effect of the treatment. By comparing the outcomes of the control group with the experimental group(s), researchers can determine whether the treatment had a significant impact.

1. Practice the data set DS\_HW1\_R\_and\_R\_data.xlsx.

Use ANOVA to analyze the factor effects of Inspector and Part. Each treatment has 3 replicates (tests). Explain the analysis results for all the following questions, including the ANOVA table and the test results.

1. Give the ANOVA table with the interaction effect.一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

   自動產生的描述
2. Are Inspector and Part the significant factors for impedance? Why?

Yes, both Inspector and Part are significant factors for impedance.

For Inspector:

* F-statistic = 38.413043
* PR (p-value) = 0.0 (approximately)

For Part:

* F-statistic = 855.642512
* PR (p-value) = 0.0 (approximately)

In both cases, the p-values are significantly smaller than 0.05, indicating that both Inspector and Part are highly significant factors affecting impedance. Therefore, they have a significant impact on impedance.

1. Is there an interaction effect between Inspector and Part to the impedance? Why?

Yes, there is a significant interaction effect between Inspector and Part on impedance.

For Inspector:Part interaction effect:

* F-value = 5.272947
* p-value (PR) ≈ 0.000001 (much less than 0.05)

The extremely small p-value (<0.05) for the interaction effect indicates that the interaction between Inspector and Part has a statistically significant impact on impedance.

1. Exam the data for model assumptions.
   1. Normality of Residuals use Shapiro-Wilk test

* Shapiro-Wilk Test Statistic: 0.8783404231071472
* Shapiro-Wilk p-value: 0.0000004913499651593

p-value<0.05 we reject the null hypothesis. In the case of the Shapiro-Wilk test, the null hypothesis is that the data follows a normal distribution.

一張含有 文字, 行, 螢幕擷取畫面, 繪圖 的圖片

自動產生的描述

* 1. Homogeneity of Variance
* Levene's Test Statistic: 0.5184678522571821
* Levene's p-value: 0.5972601919997482

p-value >0.05 we fail to reject the null hypothesis. In Levene's Test, the null hypothesis is that the variances across different groups or categories are equal (homogeneity of variances).

一張含有 文字, 螢幕擷取畫面, 行, 圖表 的圖片

自動產生的描述

一張含有 螢幕擷取畫面, 文字, 圖表, 行 的圖片

自動產生的描述

* 1. Independence

Durbin-Watson Test Statistic（DW-value）：0.485797608043757

This value is not significantly close to 2. In this case, we can consider that there is some degree of autocorrelation, but it is not very strong.

The answer is no Independence.

Code：<https://colab.research.google.com/drive/1BjLPo93FPUQDUv8yYoGVqqjgQrWorkiy?usp=sharing>