# Science Research in Qlik

## The IPC Global Methodology

**Dr. Priscila Rubim**
Data Scientist

priscila.rubim@ipc-global.com

**Igor Alcantara**
Director of Data Science & Education

igor.alcantara@ipc-global.com

## 1. Introduction

The first step in a scientific study is the formulation of a research question or hypothesis. This initial step is crucial because it guides the entire subsequent investigation process.
Formulating this question involves identifying a problem or research question, conducting a literature review, and then formulating the hypothesis. Therefore, it is necessary to verify if the intended investigation process is supported by certain pillars.
The fundamental pillars of science are concepts that ensure the integrity, validity, and credibility of scientific discoveries. Three of the most important pillars are[12]:

- **Falseability**, or refutability, is the ability of a hypothesis or theory to be tested and potentially refuted by observations or experiments. If a hypothesis cannot, in principle, be falsified, it is not considered scientific[3].
- **Replicability** is the ability of a study to be repeated, with the same methodology, and obtain the same results. It is an essential component of scientific validity[4].
- **Generality** refers to the extent to which the results of a study can be applied to different situations, populations, or contexts beyond those originally studied[5].

Replication is the general empirical mechanism for testing and falsifying theory[3]. A fundamental pillar of science that can be strongly supported using Qlik in scientific investigation. Qlik facilitates replication by enabling reproducible analyses, verification of results with different datasets, transparency and documentation, the ability to handle large volumes of data, and the identification and control of confounding variables. With its advanced data analysis capabilities, Qlik significantly supports the replication process, ensuring that analyses are robust, verifiable, and applicable to various contexts, thereby strengthening the validity and confidence in scientific discoveries.

## 2. IPC Methodology

### 2.1 Research Protocol:

A well-defined research question or hypothesis is the first step in a scientific study and guides all subsequent stages of the research. From this initial step, it is necessary to define the methodology in a research protocol. In this crucial stage, the foundation is established for a structured and systematic investigation of the study data. This process ensures that the research is conducted in a consistent and reproducible manner, providing clear guidelines for data collection, analysis, and interpretation.

### 2.2 Data Collection in Qlik

The ETL (Extract, Transform, Load) process is essential in data management, especially in scientific contexts, and is fundamental for ensuring the quality, integrity, and utility of data for analysis and research[6].

The extraction phase is the first step and involves collecting data from various sources, which can include relational databases, CSV files, APIs, scientific equipment sensors, and other raw data sources. The diversity of sources is common in scientific projects, where data can be generated by different instruments and stored in various formats[7]. Efficient extraction ensures that all relevant data is captured without losses or omissions.

After extraction, the raw data is transformed to make it suitable for analysis. This is a critical step in the ETL process as it involves cleaning, normalizing, and integrating the data.

The final phase of the ETL process is loading the transformed data into a destination system, such as a centralized database, data warehouse, or analysis platform. This destination system is where the data will be stored for future queries and analysis. Efficient loading ensures that the data is readily available to scientists and analysts without significant delays.

Effective data collection is essential for meaningful analysis and insights[8]. First, it is necessary to connect to data sources. The Data Manager or Data Load Editor can be used to establish connections with your data sources. Configure the necessary credentials and connection settings for secure access.

Once this is established, the next step is to load the data by selecting the necessary tables and fields from your data sources. Then, you can perform the required data transformations using the Data Load Editor, cleaning, aggregating, and structuring the data to meet analytical needs.

Performing ETL tasks using Qlik script language, or any script language requires knowledge of the specific language and also knowledge of data modeling techniques. Qlik platform offers a no-code alternative to that approach with Qlik Talend. Qlik Talend allows data engineers to load, transform and load the data using pipelines without the need of learning any script or modeling language.

Finally, it is crucial to validate the data to ensure accuracy and consistency[9]. Check for any anomalies or errors that need to be corrected. Qlik Talend platform offers “Data Stewardship”, a tool designed for data quality and health, where bad data entries can be automatically detected and manually fixed.

### 2.3 Using Qlik to Answer Scientific Questions

Qlik is a powerful data analytics platform that can be used to answer scientific questions through a series of systematic steps, including descriptive analysis, hypothesis testing, modeling, and visualization.

Qlik’s suite of tools facilitates a comprehensive approach to answering research questions with data. By combining data visualization capabilities with robust statistical analysis and modeling, Qlik enables researchers to gain significant insights and make data-driven decisions. Whether through detailed

descriptive analysis, rigorous hypothesis testing, or sophisticated modeling, Qlik ensures that every step of the scientific research process is supported by reliable and insightful data analysis.

- Descriptive Analysis

Use charts such as histograms, which are useful for showing the frequency distribution of a single variable, and boxplots, which are ideal for summarizing the distribution of a dataset and identifying outliers, in Qlik to visualize and describe the distribution and characteristics of your data.

- Descriptive Statistics

Calculating and displaying key descriptive statistics is essential for effectively summarizing your data. The mean represents the average value of the dataset, providing an overall view of the data's center. The median is the middle value that separates the higher half from the lower half, useful for understanding the data distribution and identifying potential skewness. The mode indicates the most frequently occurring value in the dataset, helping to identify common trends. The standard deviation measures the amount of variation or dispersion in the dataset, giving insight into the data's consistency. Together, these statistics provide a comprehensive and detailed overview of your data's characteristics, facilitating analysis and interpretation.

- Hypothesis Testing

Hypothesis testing is a critical component of data analysis, allowing researchers to make inferences and draw conclusions from their data[10]. In Qlik, you can set up and conduct statistical tests, such as t-tests and chi-2, directly within the platform. A t-test is used to determine if there is a significant difference between the means of two groups, making it ideal for comparing two distinct conditions or treatments[11]. On the other hand, chi-2 is used to compare the frequency of which an event occur between different categorical data[12]. By utilizing these tests in Qlik, researchers can rigorously assess their hypotheses and validate their findings with statistical evidence.

- Interpretation of Results

Analyzing the output of hypothesis tests involves focusing on key metrics like p-values and confidence intervals to interpret the results accurately. The p-value indicates the probability that the observed results occurred by chance, with a p-value less than 0.05 typically suggesting statistical significance[13]. This means that there is strong evidence against the null hypothesis, supporting the conclusion that the observed effect is real. Confidence intervals, on the other hand, provide a range of values that are likely to contain the true effect size, indicating the precision of the estimate. A narrow confidence interval suggests

a more precise estimate of the effect size, while a wider interval indicates less precision. By carefully examining these metrics, researchers can draw more reliable and nuanced conclusions from their data.

- Statistical Modeling Techniques

Applying advanced statistical modeling techniques within Qlik allows researchers to uncover deeper insights from their data. Linear regression is used to model the relationship between a dependent variable and one or more independent variables, enabling the prediction of outcomes based on input variables and the identification of key influencing factors[14]. This technique enhances the analytical capabilities within Qlik, allowing for more sophisticated data analysis, prediction, and decision-making processes, ultimately leading to more informed and actionable insights.

The tool available within the Qlik Platform for statistical modeling and machine learning is Qlik AutoML. Qlik AutoML provides data preprocessing tasks, such as imputation of nulls, categorical encoding, cross-validation and more. With Qlik AutoML, data scientists can train a machine learning model against different algorithms, evaluate the scores of each model and select the best model for the research in question.

Once a model is selected, it should be deployed, which means it becomes a usable asset within the Qlik Cloud platform. Researches than can access the deployed model to input their data and get in return three main outcomes:

- The predicted value
- The probability of the predicted value to occur
- The feature importance for the model

The feature importance provided is based on SHAP values, returned for each prediction (row). SHAP values evaluate the importance of each feature or model variable by analyzing how much the removal of that feature adds to the model error. The largest the error, the largest the importance.

- Model Evaluation:

Evaluating the performance and fit of your models is a crucial step in the analytical process, ensuring that the models accurately represent the data and can reliably predict outcomes. One key metric used for this purpose is the $R^2$ (Coefficient of Determination). $R^2$ indicates the proportion of the variance in the dependent variable that is predictable from the independent variables[16]. A higher $R^2$ value signifies that a greater proportion of the variance is accounted for by the model, suggesting a better fit. This metric helps researchers understand the strength and effectiveness of their models, guiding adjustments and improvements to enhance predictive accuracy and reliability. By rigorously evaluating model performance, analysts can ensure that their conclusions and predictions are based on robust and well-fitting models.

- Distributions:

Distributions are important to understand the nature of a population of sample data[20]. Qlik provides embedded functions for the most common distributions: Normal, Student's T, Beta, Binomial, Chi2, F, Gamma, and Poisson. Additionally, with Qlik Script, researchers can implement additional functionalities to the distributions such as creating simulations like the Montecarlo Simulation.

A Distribution function in Qlik can be used to plot distributions charts or to calculate distribution specific values like the probability of a value, the value given a probability or the accumulated probability. Qlik also have functions to calculate the skewness and kurtosis of a distribution.

- Correlation:

Correlation is a very important tool for one working with research[17]. Qlik provides two functions related to correlation: *Correl*, which gives the Pearson $R^2$ correlation score, and *MutualInfo*, which gives the mutual information between two fields or between two aggregated values. Mutual Information Analysis quantifies the amount of information obtain about one random variable through another random variable, measuring dependency between them[18]. MutualInfo function allows three types of analysis: pair-wise mutual information, driver breakdown by value, and feature selection[19].

- Visualization and Reporting

Visualization and reporting are key components of data analysis, transforming complex data into easily interpretable insights. In Qlik, creating interactive and dynamic dashboards is an effective way to present your findings. These dashboards allow users to explore the data through various filters and selections, enabling a deeper and more nuanced understanding of the results. Interactive elements, such as clickable charts and real-time data updates, empower users to drill down into specific details and discover trends or anomalies that might not be immediately apparent. By providing a flexible and engaging way to interact with the data, Qlik dashboards enhance the ability to communicate insights clearly and effectively, facilitating better decision-making and collaboration among stakeholders.

Qlik empowers researchers to conduct thorough and effective scientific inquiries, from initial data exploration to final reporting, ensuring that every research question is answered with precision and clarity.

## 3. Conclusion

Using Qlik to answer scientific questions is a comprehensive approach that combines descriptive analysis, hypothesis testing, statistical modeling, and interactive visualization. Descriptive analysis in Qlik allows for the calculation and display of essential statistics such as mean, median, mode, and standard deviation, providing a detailed view of the data. Hypothesis tests, such as t-tests and chi-2, can be set up and conducted directly in Qlik, allowing for the interpretation of results through p-values and confidence intervals.

Advanced statistical modeling techniques, such as linear and logistic regression, help uncover deeper insights and predict outcomes with precision. Model evaluation is crucial to ensure accuracy and usefulness, with metrics like $R^2$ indicating the proportion of variance explained by the models.

Furthermore, creating interactive dashboards in Qlik enables effective presentation of results. These dashboards facilitate data exploration through filters and selections, offering a deeper and more detailed understanding of the results. Interactive elements enhance the communication of insights, promoting better decision-making and collaboration among stakeholders.

In summary, Qlik is a versatile and highly useful tool that enables researchers to conduct thorough and effective scientific investigations, from initial data exploration to final reporting, ensuring that each research question is answered with precision and clarity.

## 4. Appendix: Images


*Figure 1: The IPC Global Methodology Framework*

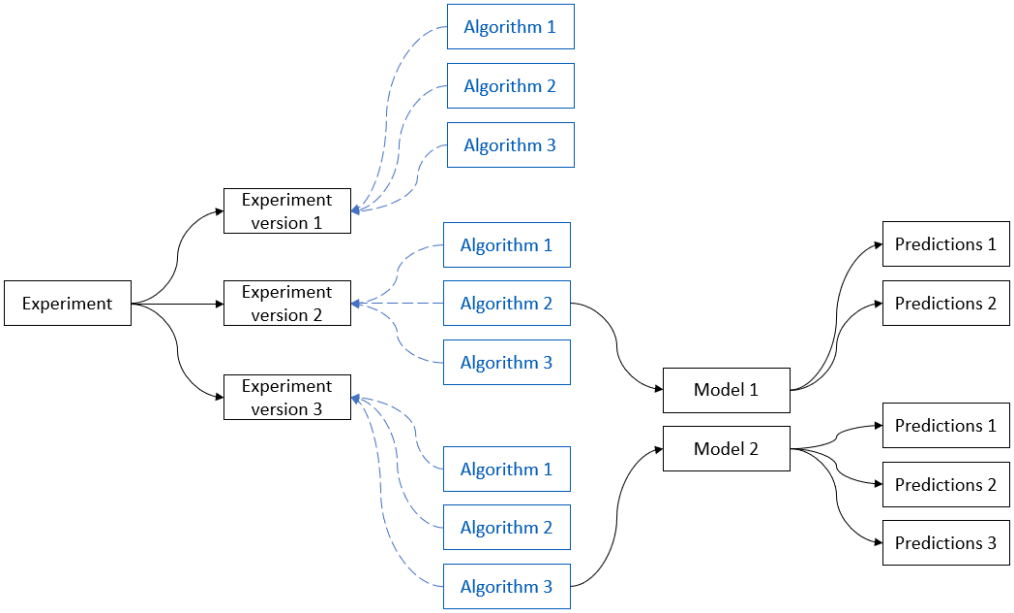


*Figure 2: Qlik AutoML workflow*

What is the effect of Remdesivir on a patient's heart rate?

**Null Hypothesis:** Heart Rate is the same regardless of the use of Remdesivir.

with **2,184** degrees of freedom, the t-value is **13.57**. The t-table reference value is **1.98**.
Therefore , we **Reject the Null Hypothesis**.
The estimated p-value is **~0.005**. The significance value was set to **0.05**.

*Figure 3: Example of Hypothesis Test in Qlik*

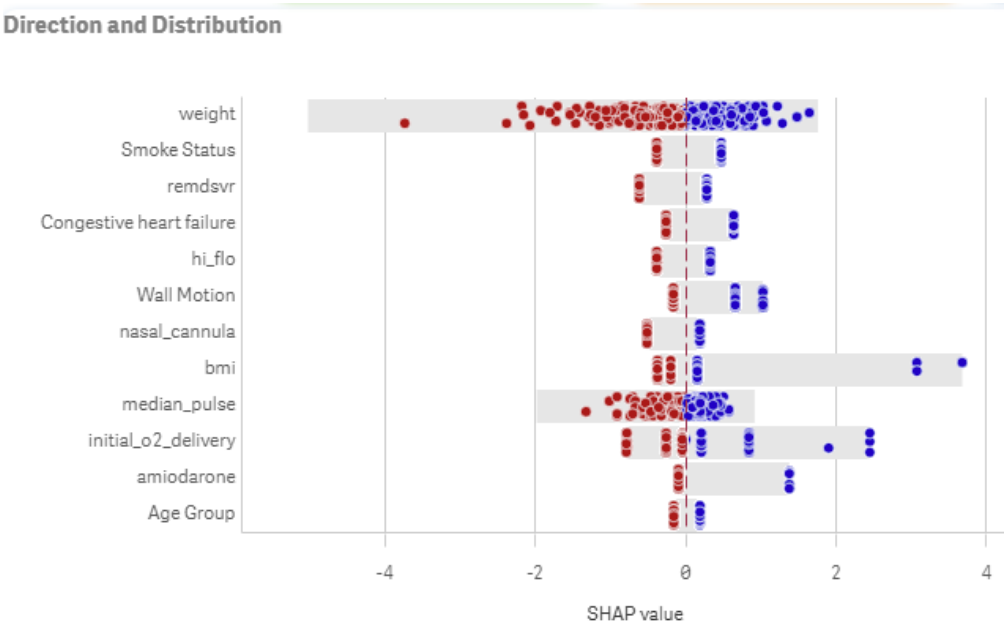


*Figure 4: Example of SHAP Values in Qlik*

Characteristics of target column that determine model type

| Model type | Number of distinct values in column | Feature type required | Additional information |
|---|---|---|---|
| Binary classification | 2 | Any | - |
| Multiclass classification | 3-10 | Any | A column with more than 10 distinct, non-numeric classes is not selectable as the target. |
| Regression | More than 10 | Numeric | - |

*Figure 5: Types of models in Qlik AutoML*

## 5. References


1. LeBel, E., Berger, D., Campbell, L., & Loving, T. (2017). Falsifiability is not optional.. Journal of personality and social psychology, 113 2, 254-261 . https://doi.org/10.1037/pspi0000106.

2. LeBel, E., McCarthy, R., Earp, B., Elson, M., & Vanpaemel, W. (2018). A Unified Framework to Quantify the Credibility of Scientific Findings. Advances in Methods and Practices in Psychological Science, 1, 389 - 402. https://doi.org/10.1177/2515245918787489.

3. (2017). "Falsifiability is not optional": Correction to LeBel et al. (2017).. *Journal of personality and social psychology*, 113 5, 696 . https://doi.org/10.1037/pspi0000117.

4. Nosek, B., Hardwicke, T., Moshontz, H., Allard, A., Corker, K., Dreber, A., Fidler, F., Hilgard, J., Struhl, M., Nuijten, M., Rohrer, J., Romero, F., Scheel, A., Scherer, L., Schönbrodt, F., & Vazire, S. (2020). Replicability, Robustness, and Reproducibility in Psychological Science. Annual review of psychology. https://doi.org/10.31234/OSF.IO/KSFVQ.

5. Green, L., & Glasgow, R. (2006). Evaluating the Relevance, Generalization, and Applicability of Research. *Evaluation & the Health Professions*, 29, 126 - 153. https://doi.org/10.1177/0163278705284445.

6. Homayouni, H. (2018). Testing Extract-Transform-Load Process in Data Warehouse Systems. *2018 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*, 158-161. https://doi.org/10.1109/ISSREW.2018.000-6.

7. Humphrey, M., Agarwal, D., & Ingen, C. (2009). Fluxdata.org: Publication and Curation of Shared Scientific Climate and Earth Sciences Data. *2009 Fifth IEEE International Conference on e-Science*, 118-125. https://doi.org/10.1109/E-SCIENCE.2009.25.

8. Dikilitaş, K., & Griffiths, C. (2017). Collecting the Data. , 107-127. https://doi.org/10.1007/978-3-319-50739-2_5.

9. Lindholm, A., Zachariah, D., Stoica, P., & Schön, T. (2018). Data Consistency Approach to Model Validation. *IEEE Access*, 7, 59788-59796. https://doi.org/10.1109/ACCESS.2019.2915109.

10. Slowiaczek, L., Klayman, J., Sherman, S., & Skov, R. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer?. *Memory & Cognition*, 20, 392-405. https://doi.org/10.3758/BF03210923.

11. Mishra, P., Singh, U., Pandey, C., Mishra, P., & Pandey, G. (2019). Application of Student's t-test, Analysis of Variance, and Covariance. *Annals of Cardiac Anaesthesia*, 22, 407 - 411. https://doi.org/10.4103/aca.ACA_94_19.

12. Larson, M. (2008). Analysis of Variance. Circulation, 117, 115-121. https://doi.org/10.1161/CIRCULATIONAHA.107.654335.

13. Andrade, C. (2019). The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives. Indian Journal of Psychological Medicine, 41, 210 - 215. https://doi.org/10.4103/IJPSYM.IJPSYM_193_19.

14. Kumari, K., & Yadav, S. (2018). Linear regression analysis study. *Journal of the Practice of Cardiovascular Sciences*, 4, 33 - 36. https://doi.org/10.4103/JPCS.JPCS_8_18.

15. Puhr, R., Heinze, G., Nold, M., Lusa, L., & Geroldinger, A. (2017). Firth's logistic regression with rare events: accurate effect estimates and predictions?. *Statistics in Medicine*, 36, 2302 - 2317. https://doi.org/10.1002/sim.7273.

16. Nakagawa, S., Johnson, P., & Schielzeth, H. (2016). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14. https://doi.org/10.1098/rsif.2017.0213.

17. Asuero, A. G., Sayago, A., & González, A. G. (2006). The Correlation Coefficient: An Overview. Critical Reviews in Analytical Chemistry, 36(1), 41–59. https://doi.org/10.1080/10408340500526766

18. Veyrat-Charvillon, N., Standaert, FX. (2009). Mutual Information Analysis: How, When and Why?. In: Clavier, C., Gaj, K. (eds) Cryptographic Hardware and Embedded Systems - CHES 2009. CHES 2009. Lecture Notes in Computer Science, vol 5747. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04138-9_30

19. Vergara, J.R., Estévez, P.A. A review of feature selection methods based on mutual information. Neural Comput & Applic 24, 175–186 (2014). https://doi.org/10.1007/s00521-013-1368-0

20. Krishnamoorthy, K. (2006). Handbook of Statistical Distributions with Applications (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420011371

# About IPC Global

IPC Global is a diverse, nationwide team of experienced, engaging, and effective data analytics experts. We provide advisory, consulting and managed services to solve your organization's Enterprise Intelligence challenges.

For more than 20 years we have been the go-to source for Artificial intelligence, Data Analytics, Data Science & Research, Business intelligence, Cloud Solutions, Data Integration, Education, Advisory services and more.