**Predicting Critical Temperature by Superconductivity of data**

By: Submitted To:

Atharva Todmal Dr Dhanpratap Singh: 25706

Registration Number: 12010145 School of Computer Science and Engineering

School of Computer Science and Engineering, Lovely Professional University

Lovely Professional University

# Abstract

The aim of this article is to elucidate and exemplify the principal component regression (PCR) technique, which is a regression approach based on principal component analysis (PCA) - a dimensionality reduction technique. PCR is a regression technique commonly employed for datasets with high multicollinearity. This article commences with a brief overview of the background, objectives, and mathematical assumptions required for PCA and PCR. A practical demonstration of the PCR technique is then presented, utilizing the superconductivity dataset from the UCI machine learning repository. The demonstration covers several preliminary steps of a regression analysis, including exploratory data analysis, model building, contextual interpretation, and residual analysis, to provide appropriate context for the PCR method. The paper concludes by discussing the advantages and disadvantages of PCR, and offering future recommendations for prospective PCR model builders. It is advisable to have a sound understanding of linear algebra and multiple linear regression to grasp the full essence of this paper. All statistical analyses and data cleaning performed for this paper were conducted using Python, employing various packages such as Pandas, NumPy, Matplotlib, Seaborn, Scikit-Learn, and SciPy. Some data visualizations were generated using R statistical software, utilizing the ggplot2 and plot\_ly packages.

# Introduction

## Background

Regression is a statistical tool used for prediction and identifying relationships between variables. Its main goal is to make predictions based on a set of predictor variables. Regression is a versatile technique and is widely used in many fields of study. However, one major issue that can arise in regression analysis is multicollinearity, which is when predictor variables are highly correlated with each other. This can lead to inflated variances of the regression coefficients, which can affect the model's accuracy. To address this issue, principal component regression (PCR) is used.

PCR is an ordinary least squares regression technique that requires certain assumptions, including independent and identically distributed observations, approximately normally distributed residuals, and uncorrelated predictor variables. However, in real-world applications, multicollinearity is a common problem that can cause issues for regression model building. In such situations, PCR can help mediate high multicollinearity.

The PCR procedure involves conducting principal component analysis (PCA) on the centered design matrix of the dataset, which transforms the original set of variables into a set of linearly uncorrelated variables called principal components. The goal with PCA is to preserve as much variance as possible while reducing the dimensionality of the dataset. This reduces multicollinearity, which can help improve the accuracy of the regression model.

To perform PCR, the centered design matrix is decomposed into three matrices using the singular value decomposition (SVD) method. The principal components are then found by multiplying the centered design matrix with the right singular vectors. A subset of the principal components is used as new predictor variables for the regression model, and the coefficients of the model are found using ordinary least squares regression. The coefficients can be transformed back into the units of the original predictors, and the accuracy of the model is evaluated based on the results.

Overall, PCR is a useful technique for addressing multicollinearity in regression analysis. It is based on principal component analysis, which is a dimensionality reduction technique that preserves as much variance as possible while reducing dimensionality. This can help improve the accuracy of regression models and make them more applicable in real-world situations.

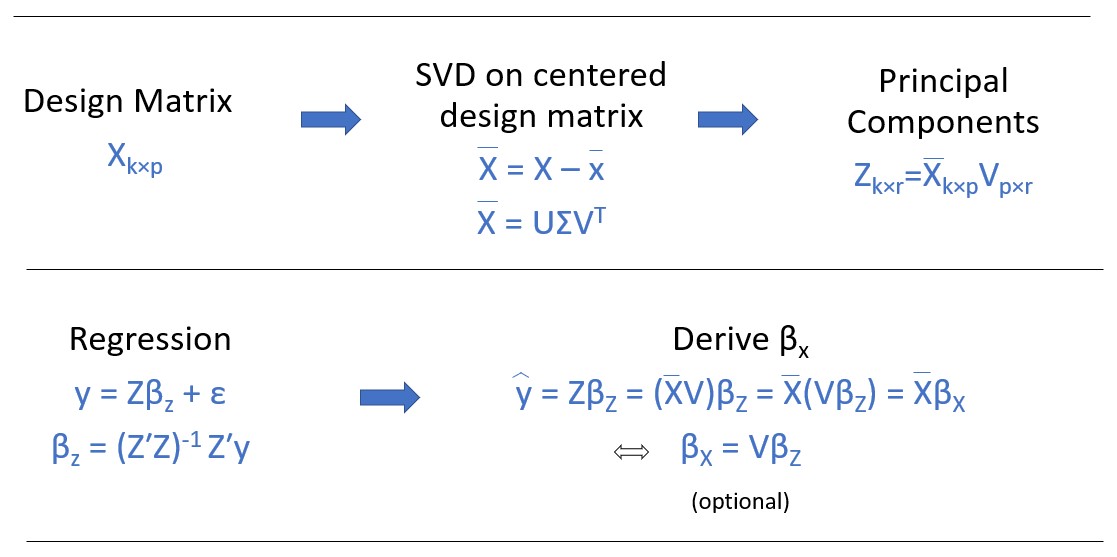


Figure 1: The graphic above shows the PCR procedure summarized in a step-by-step process using mathematical notation. The last step labeled ‘optional’ is only needed if the coefficients in the model are needed to be interpreted in the context and units of the original standardized variables.

# Methods

## Data Description and Setup

The dataset used in this analysis was obtained from the UCI machine learning repository. It is a high-dimensional dataset with mainly quantitative variables that have high multicollinearity. The dataset contains the main dataset used for prediction which has the critical temperature and 81 other variables on 21,263 superconductors, resulting in 82 columns and 21,263 rows. There is also a second dataset that contains information on the element composition of the superconductors, resulting in a dataset with 87 columns representing the elements of the periodic table hydrogen through radon and the 21,263 superconductor observations organized as rows. The goal of this analysis is to predict the critical temperature (𝑇𝑐) of a superconductor using 81 features based on properties of the elemental composition of each superconductor material. The property features were extracted from the observed superconductors. The elemental properties are summarized in Table 1. For this paper, the PCR model was built using the features dataset and excluding the elements dataset.

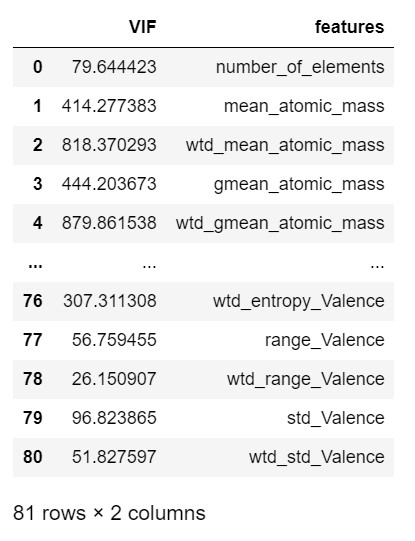
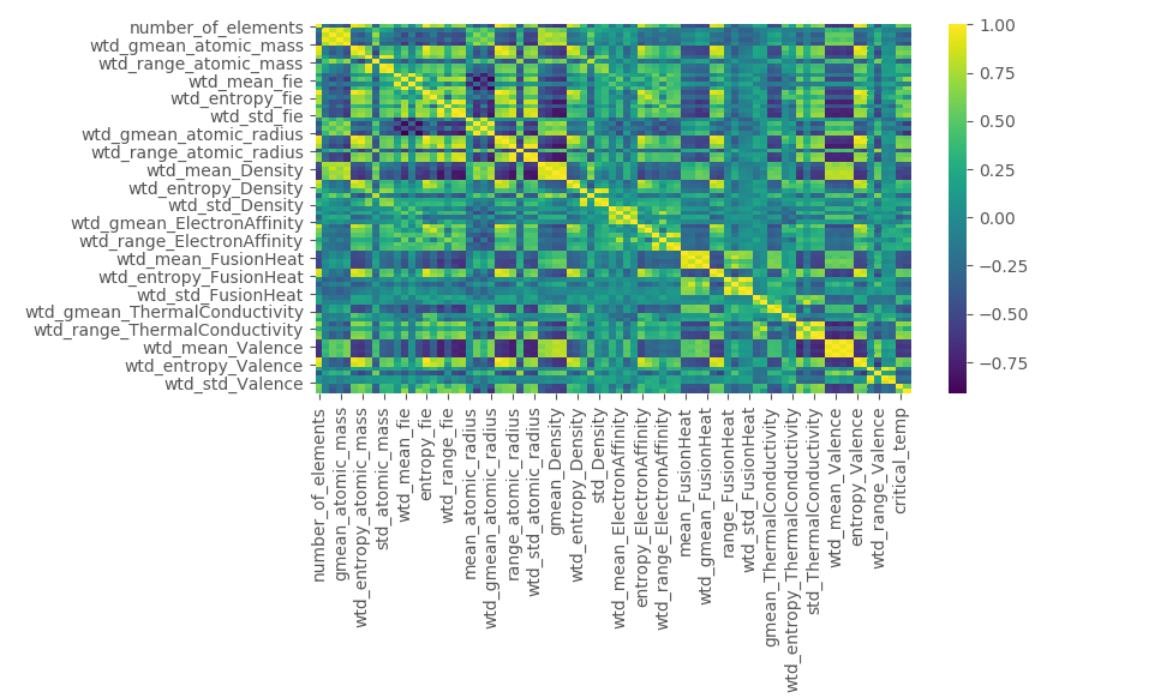
|  |  |
| --- | --- |
| **Property** | **Units** |
| Atomic Mass | Atomic mass units (AMU) |
| First Ionization Energy | Kilo-Joules per mole (kJ/mol) |
| Atomic Radius | Picometer (pm) |
| Density | Kilograms per meters cubed (kg/m3) |
| Electron Affinity | Kilo-Joules per mole (kJ/mol) |
| Fusion Heat | Kilo-Joules per mole(kJ/mol) |
| Thermal Conductivity | Watts per meter-Kelvin (W/(m K)) |
| Valence | No units (integer) |

Table 1: The table shows the main elemental property in the dataset. The table excludes the property features extracted from each superconductor (used as predictors), which include geometric mean, range, and standard deviation. A more detail explanation of the variables can be found in the relevant scientific paper of the dataset.

## Exploratory Data Analysis

A major point for the exploratory data analysis section of the dataset was to detect multicollinearity in the data. Graphical representations of multicollinearity are shown in the form of a correlation matrix heatmap (figure 2) and a variance inflation factor (VIF) table (table 2).

Table 2: VIF table of the predictor variables shows extremely high multicollinearity for multiple predictor Figure 2: Correlation matrix heatmap of the data shows variables. very high correlation between several variables.



The correlation matrix heatmap shows very high correlation between several variables such as the geometric mean and weighted mean features for atomic mass. The VIF table shows extremely high multicollinearity between many predictors, since virtually all features had a VIF much higher than 10. This shows the degree of multicollinearity in the data which assuredly would produce an ordinary least squares regression model that cannot be reliably interpreted in any significant way. The distribution of the response variable can be observed in the histogram constructed below (figure 3).

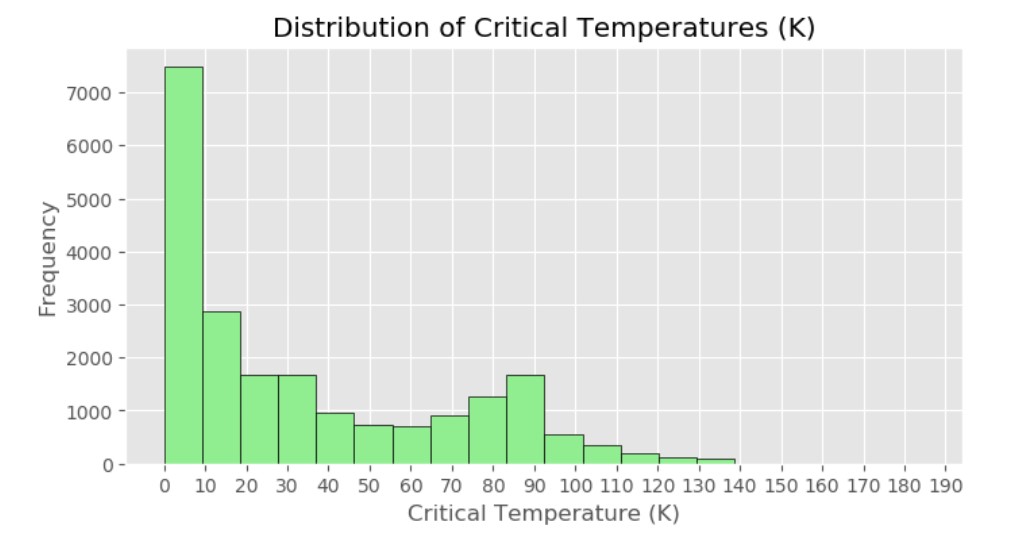


Figure 3: The histogram roughly shows the distribution of the response variable, critical temperature. The distribution is clearly right skewed since most of the critical temperatures are in between 0 K and 10 K.

The distribution of critical temperatures is clearly right skewed, indicating that most superconductors have a critical temperature that falls in between 0 K and 10 K. Furthermore, scatterplots between each predictor variable and critical temperature were made to investigate the individual relationships in the data, found in the appendix (figure 4). Many of the individual scatterplots did not clearly show a linear relationship which suggests that transformations on either the response variables or any of the predictor variables could be necessary in a full regression analysis. PCA score plots were also made to explore differences between superconductors with and without certain elements, where each element was column bind to the principal component matrix as a factor variable to make the plots. Several elements were examined but oxygen was of particular interest because the PCA score plot showed a considerable amount of separation between superconductors with oxygen and those without. This motivated building a second PCR model using oxygen content as an indicator variable. The PCA score plots for oxygen can be found in the appendix (figures 5, 6).

## Model Building

The PCR model was built by doing PCA on the standardized design matrix, as opposed to the centered design matrix. This can be beneficial because standardization eliminates the influence of wildly different units on the principal components. Subsetting the principal components should be done with preserving as much variance as possible, which can be determined in a few different ways. One method is to calculate the cumulative explained variance by each subsequent principal component. The

𝜆𝑚 *th* explained variance for each principal component is found by ∑ 𝜆𝑖 where 𝜆𝑚 represents the *m* eigenvalue of the correlation matrix of the standardized design matrix. Another method is to keep only the principal components who have a large corresponding eigenvalue (usually those with 𝜆𝑚 > 1). A cumulative explained variance plot and a scree plot can both be found in the appendix of this paper (figures 7, 8). These methods show that approximately 99% of the variance in the original dataset can be explained by the first 30 principal components. Consequently, the PCR model was built by fitting the first 30 principal components against the response 𝑇𝑐 (model 1).

To explore using categorical variables with PCR, oxygen content was used as an indicator variable. To do this, the data was organized so that the features of all 9299 superconductors that did not contain oxygen were stacked on top of the features of all 11964 superconductors that did contain oxygen. PCA was then done to this data matrix. Then a new factorized variable was created where each superconductor that contained no oxygen was assigned a ‘0’ while each superconductor that had any recorded amount of oxygen in them was assigned a ‘1’. This factorized variable was then combined with the principal components accordingly (after PCA was done on the quantitative variables) to build a second PCR model by fitting the first 30 principal components with the indicator variable oxygen against the response 𝑇𝑐 (model 2).

Generalized cross-validation (10-fold cross-validation) measures, including the coefficient of determination (R2) and mean squared error (MSE), were used as measures to assess how well each model predicted the observed values. The appendix also contains a graphical display for each regression model to help visualize the predictive performance of each model (figures 9, 12) and residual plots for each model to see if regression residual assumptions were satisfied (figures 10, 11, 13, 14).

# Results

The cross-validation summary results for each model are displayed below (table 3). Model 1 produced an MSE of 393.389 and an R2 of 0.665. Model 2 produced an MSE of 403.463 and an R2 of 0.656. The cross-validation measures for both models were quite similar and are not terrible results. Although not great, the cross-validation measures indicate two decent models for predictions. A visual display of the regression lines for model 2 can be seen on the next page (figure 12).

|  |  |  |
| --- | --- | --- |
| **Metrics** | **Model 1** | **Model 2** |
| MSE | 393.386 | 403.463 |
| R2 | 0.665 | 0.656 |

Table 3: The table shows the cross-validation results for each model. Model 1 refers to fitting the response 𝑇𝑐 against the first 30 principal components. Model 2 refers to fitting the response 𝑇𝑐 against the first 30 principal components and with an indicator variable for oxygen content

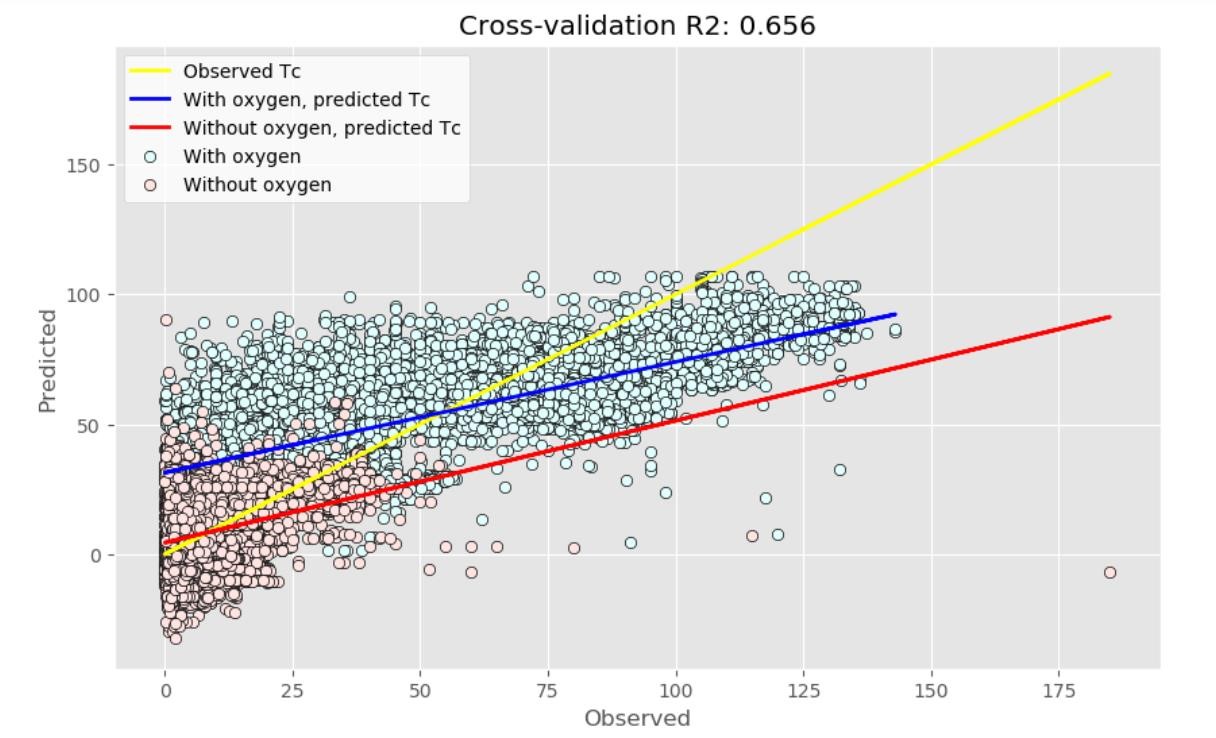


Figure 12: The above scatterplot shows a visual display of the regression lines for model 2. The yellow line represents the observed 𝑇𝑐 values plotted against each other. The blue line represents the predicted 𝑇𝑐 values for superconductors with oxygen while the red line represents the predicted 𝑇𝑐 values for superconductors without oxygen. The blue dots represent predicted 𝑇𝑐 values for superconductors with oxygen while the red dots represent predicted 𝑇𝑐 values for superconductors without oxygen, both plotted against their corresponding observed 𝑇𝑐 values.

# Discussion

## The Regression Models

The results of the cross-validation for the two models were not impressive, but they showed promise. PCR is essentially a regression problem and there are various methods to enhance the model, such as applying transformations to either the response or predictors and performing a residual analysis to identify high influence or leverage points. Additionally, in this analysis, the entire elements dataset was not included in the model building phase, except for oxygen. If more information about the element composition was added, it is likely that the predictive power of the PCR model would increase.

PCR has both advantages and disadvantages. One of the key advantages of PCA is its ability to reduce the dimensions of the data, which is extremely beneficial in regression. One of the most frequently applied reasons for using PCR is to address high multicollinearity in the data. This approach works well when the researcher wants to incorporate all available predictor information, precluding the use of variable selection algorithms. Another advantage of PCR is the data visualization features of PCA. PCA is an excellent tool for data visualization because it allows for reduced dimensional space visualization, making it easier to interpret high dimension datasets. This is particularly useful in identifying group differences and patterns in the data, as demonstrated by the PCA score plots for oxygen (figures 5, 6). In addition, PCA loading plots can help to determine the influence of specific predictor variables on the principal components.

PCR does have a few drawbacks. The model building process is slightly more complex than most regression techniques, but with a better understanding of linear algebra, it is relatively simple and straightforward. Another disadvantage is the difficulty of analyzing categorical variables, since PCA works with a centered data matrix and it doesn't make computational sense to center an indicator variable. In this example, the principal component entries were grouped by an indicator variable after applying the PCA technique to the quantitative variables.

The main disadvantage of PCR is the difficulty in interpreting the model itself. Although it can be used for predictions, interpreting the coefficients of the model in the context of the original units and predictor variables requires additional transformations. Furthermore, PCA is often called a 'plug-and-play' technique because any dataset can be input, and output is guaranteed, regardless of the context of the original dataset. This can be seen as a disadvantage because the response variable is completely excluded from the PCA. However, it can also be seen as an advantage because it is a non-parametric technique, and the output is unique and independent of the user.

# Conclusion

PCR is a valuable statistical tool for constructing regression models that involve high multicollinearity. It is often used in conjunction with other models such as ridge regression or lasso regression to compare different models. PCR can also be used to simplify and visualize complex data. However, further research is needed to determine exactly how and when multicollinearity is reduced. Most textbooks and online sources suggest that the reduction of multicollinearity can be achieved by using a subset of principal components in regression analysis. However, none of these sources specify the optimal size of the subset. For example, in this study, a subset of 30 principal components out of 81 total components was used to build the regression model, effectively reducing multicollinearity by removing over 50 principal components. However, the paper did not mention any specific technique for measuring multicollinearity after subsetting the principal components. As a result, further research is required to fully understand the effects of PCR on multicollinearity reduction.

# References

“3D Scatter Plots in R”, plotly Graphing Libraries, 2020, <https://plotly.com/r/3d-scatter-plots/>

Alice, Michy, “Performing Principal Components Regression (PCR) in R”, Milano R, 2016, <http://www.milanor.net/blog/performing-principal-components-regression-pcr-in-r/>

Bremer, Martina, “Multicollinearity,” Lecture 13, Math 261A, San Jose State University, 2018

## Hamidieh, Kam. A data-driven statistical model for predicting the critical temperature of a superconductor, Computational Materials Science, Volume 154, November 2018, Pages 346-354

Johnson, Richard, and Dean Wichern. *Applied Multivariate Statistical Analysis,* sixth edition, Pearson Education Inc., 2013.

## Kanyongo, Gibbs, Determining the Correct Number of Principal Components to Extract from a Principal

*Components Analysis: A Monte Carlo Study of the Accuracy of the Scree Plot*, Journal of Modern Applied Statistical Methods, Vol. 4 Issue 1, Article 13, 2005,

<https://digitalcommons.wayne.edu/cgi/viewcontent.cgi?article=1166&context=jmasm>

Montgomery, Douglas, and Elizabeth Peck, Geoffrey Vining. *Introduction to Linear Regression Analysis*, fifth edition, John Wiley & Sons Inc., 2012

Pelliccia, D. (2019, September 14). Principal Components Regression vs Ridge Regression on NIR data in Python. <https://nirpyresearch.com/pcr-vs-ridge-regression-nir-data-python/>

Piccolo, Brian, “Plotting PCA Scores and Loadings in Scatterplots,” RPubs by RStudio, 2017, <https://www.rpubs.com/bpiccolo/pcaplots>

“Principal Component Regression”, Wikipedia the Free Encyclopedia, 2020, <https://en.wikipedia.org/wiki/Principal_component_regression>

Shimizu, K., and K. Suhara, M. Ikumo, M. I. Eremets and K. Amaya. “Superconductivity in Oxygen”,

Nature, 393, 767-769, 1998, <https://doi.org/10.1038/31656>

Sobolewska, Ewa, “Principal Component Regression,” RPubs by RStudio, 2019, <https://rpubs.com/esobolewska/pcr-step-by-step>

“Superconductors and Superconducting Materials Information”, Engineering 360, IEEE GlobalSpec, 2020. [https://www.globalspec.com/learnmore/materials\_chemicals\_adhesives/electrical\_optical\_specialty\_m aterials/superconductors\_superconducting\_materials](https://www.globalspec.com/learnmore/materials_chemicals_adhesives/electrical_optical_specialty_materials/superconductors_superconducting_materials)

“Superconductivity Data Data Set”, UCI Machine Learning Repository, 2018, <https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data>

Suryanarayana, T.M.V., and P. B Mistry. *Principal Component Regression for Crop Yield Estimation*, Springer Singapore, 2016.

# APPENDIX

## Contains all figure/tables used in paper and Python/RStudio code used for analysis

Figure 1:

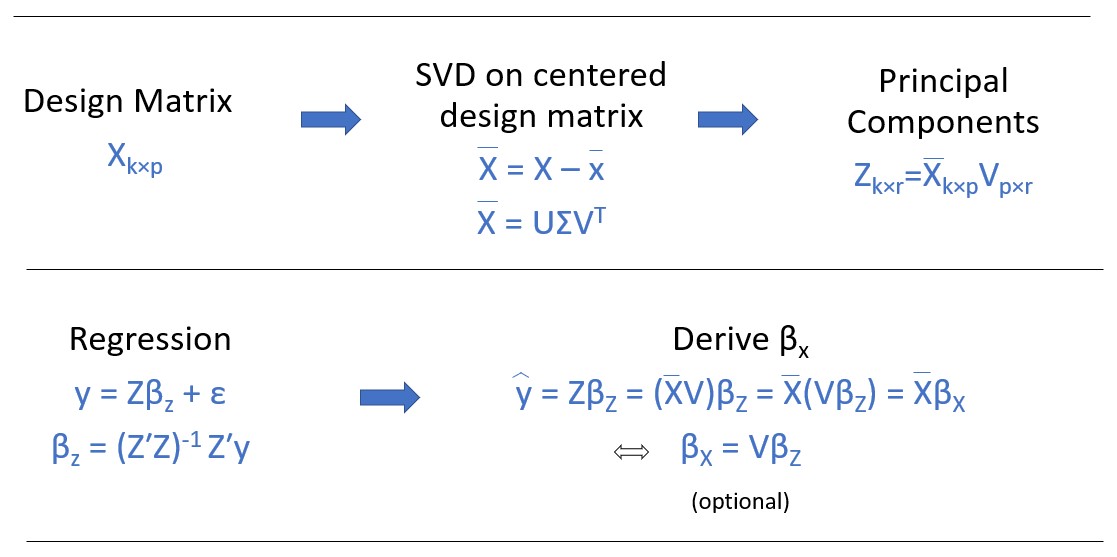


Figure 1: The graphic above shows the PCR procedure summarized in a step-by-step process using mathematical notation. The last step labeled ‘optional’ is only needed if the coefficients in the model are needed to be interpreted in the context and units of the original standardized variables (PowerPoint).

Table 1:

|  |  |
| --- | --- |
| **Property** | **Units** |
| Atomic Mass | Atomic mass units (AMU) |
| First Ionization Energy | Kilo-Joules per mole (kJ/mol) |
| Atomic Radius | Picometer (pm) |
| Density | Kilograms per meters cubed (kg/m3) |
| Electron Affinity | Kilo-Joules per mole (kJ/mol) |
| Fusion Heat | Kilo-Joules per mole(kJ/mol) |
| Thermal Conductivity | Watts per meter-Kelvin (W/(m K)) |
| Valence | No units (integer) |

Table 1: The table shows the main elemental property in the dataset. The table excludes the property features extracted from each superconductor (used as predictors), which include geometric mean, range, and standard deviation. A more detail explanation of the variables can be found in the relevant scientific paper of the dataset (Word).

Figure 2:

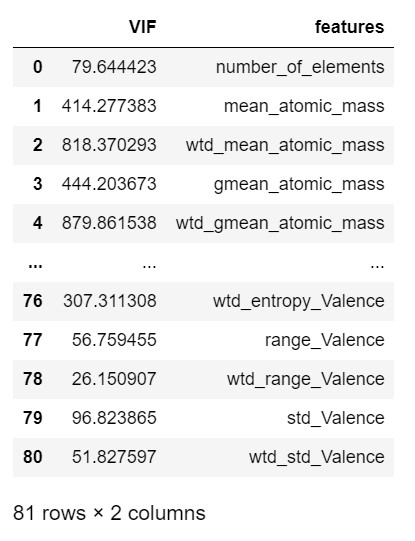


Figure 2: Correlation matrix heatmap of the data shows very high correlation between several variables. For instance, there is a high correlation (0.848) between the geometric mean and weighted mean features for atomic mass (Python).

Table 2:

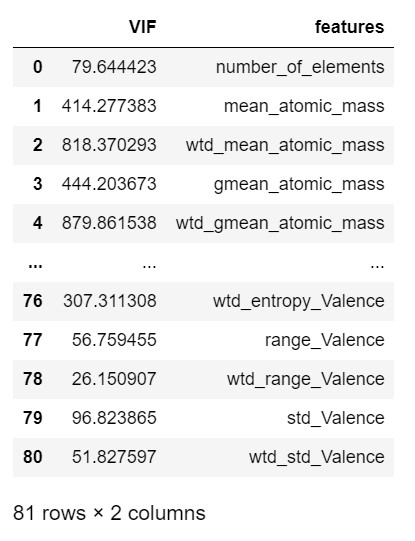


Table 2: VIF table of the predictor variables shows extremely high multicollinearity for multiple predictor variables. A general rule-of-thumb is that a VIF higher than 10 shows high multicollinearity in that predictor variable. Here we see nearly all predictors having extremely high

multicollinearity. For instance, the feature

*wtd\_mean\_atomic\_mass* has a VIF of approximately 818.37 (Python).

Figure 3:

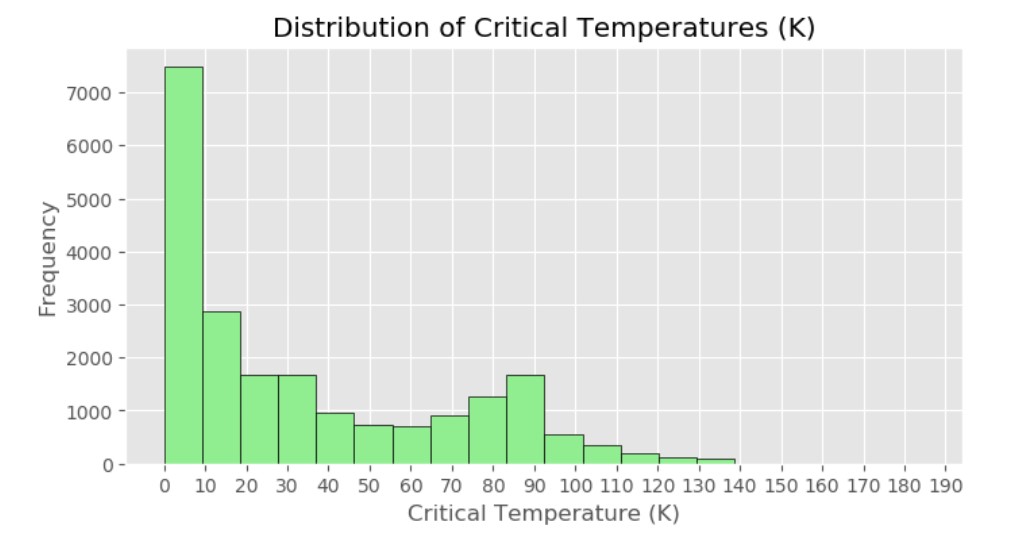


Figure 3: The histogram roughly shows the distribution of the response variable, critical temperature. The distribution is clearly right skewed since most of the critical temperatures are in between 0 K and 10 K. In theory, a conductor becomes superconductive after it drops below a certain temperature, called the critical temperature. This temperature is usually very cold which is why the critical temperature is usually measured in Kelvin. A superconductor is defined as a material that has no resistance to electrical current (Python).

Figure 4:

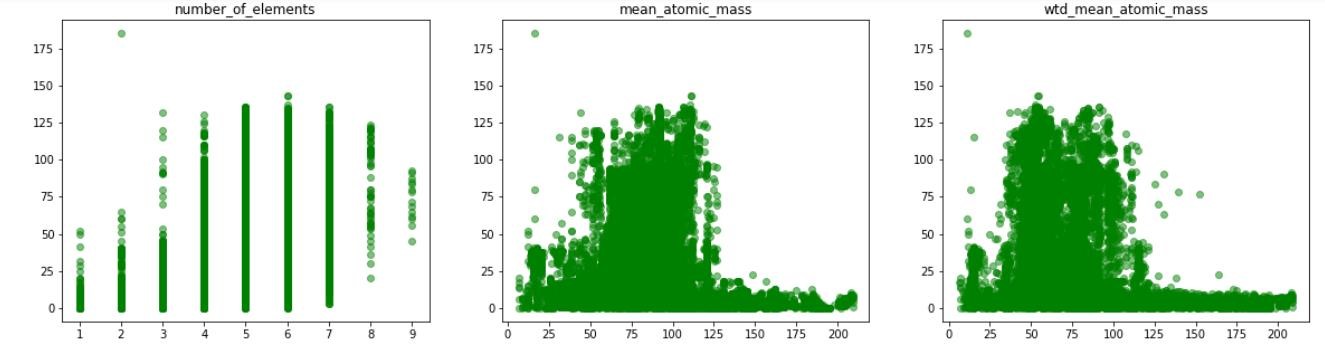


Figure 4: Here there are examples of three scatterplots made during the exploratory data analysis portion of the paper. They show critical temperature plotted against three predictor variables, *number\_of\_elements*, *mean\_atomic\_mass*, and *wtd\_mean\_atomic\_mass*. There appears to be a weak linear relationship between the response and predictors, indicating that transformations on either the response or predictors should be explored in the model building process. Scatterplots like these were constructed for every predictor variable to briefly observe the relationship between each response and its predictors (Python).

Figure 5:

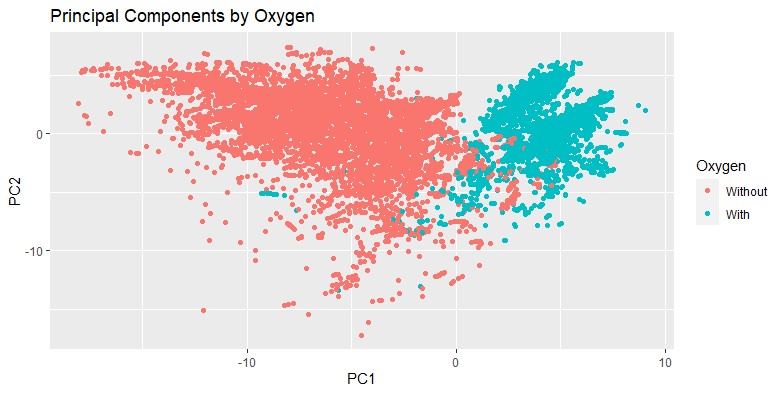


Figure 5: This is referred to as a score plot and it shows first two principal components plotted against each other, where the observations are grouped by superconductors with oxygen and those without oxygen (after PCA was calculated on the continuous variables). The score plot shows clear separation between the groups along the first principal component axis, which accounts for about 39% about the explained variance. Furthermore, there seems to be less variability within the oxygen group than within the no oxygen content group. There appears to be much less separation across the second principal component (RStudio).

Figure 6:

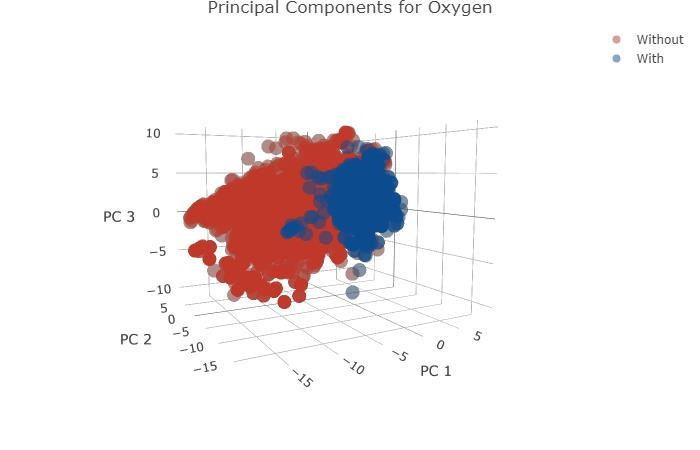


Figure 6: This plot depicts a 3D scatterplot of the first 3 principal components plotted against each other, which is an extension of figure 5 by using a third dimension. Again, we see clear

separation between superconductors with oxygen and those without. The superconductors with oxygen seem to be much more similar to each other than those without oxygen (RStudio).

Figure 7:

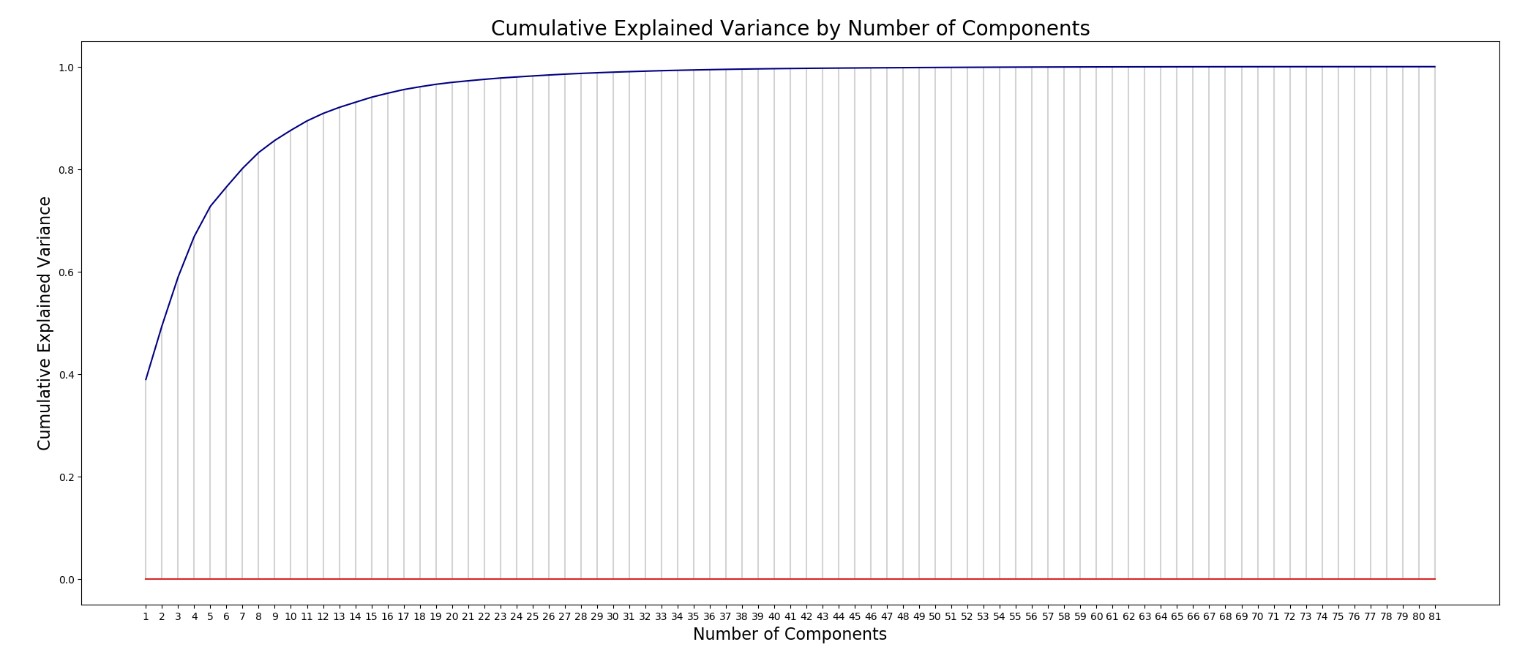


Figure 7: The above plot shows the cumulative explained variance by each subsequent principal component. The first principal component has already explained approximately 39% of the variance in the whole dataset. By 30 principal components, approximately 99% of the variance in the dataset is explained. This was the criteria used to decide on a subset of principal components to be used in the PCR model (Python).

Figure 8:

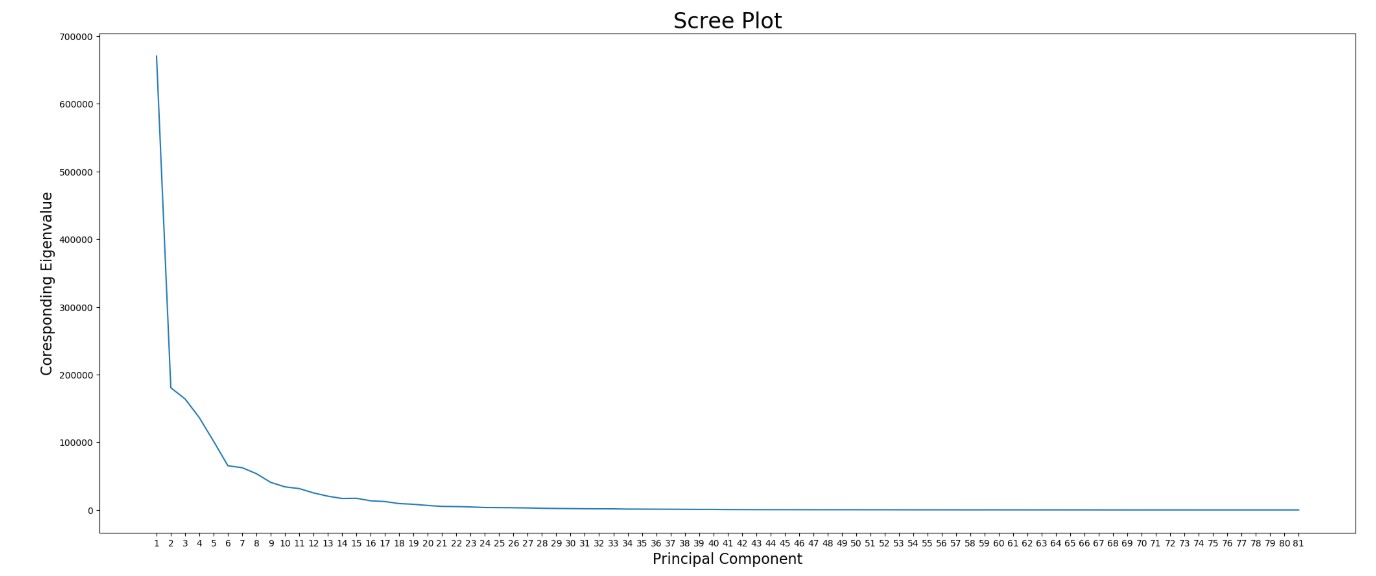


Figure 8: The above plot shows a scree plot of the principal components. This plot shows the corresponding eigenvalue plotted against its corresponding principal component. This is another method used to determine the number of principal components needed in the subset. The goal here would be to only keep the principal components that contain a ‘large’ enough eigenvalue (usually greater than 1). Although, in this example every principal component had an eigenvalue greater than 1. The explained cumulative variance method was used instead to attempt to subset the principal components appropriately (Python).

Figure 9:

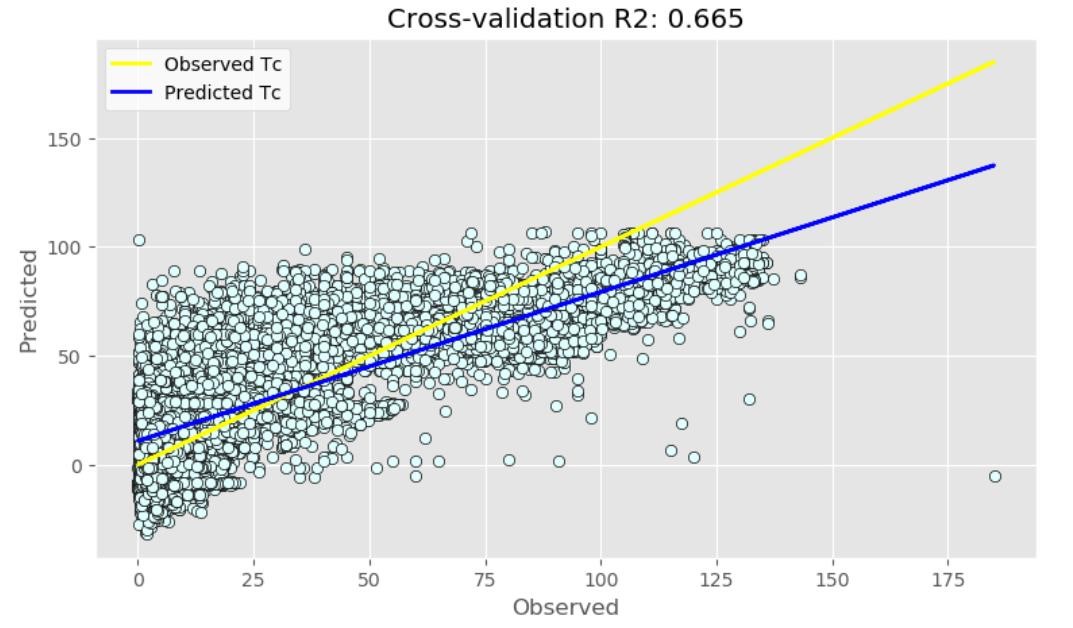
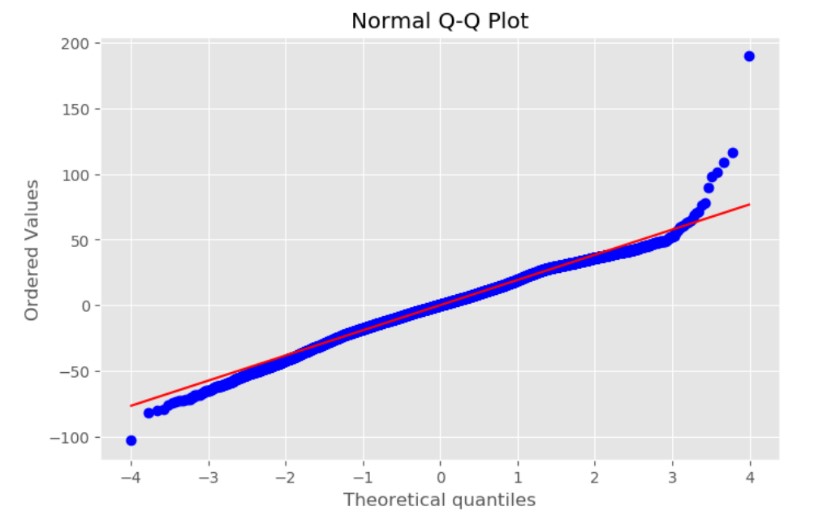
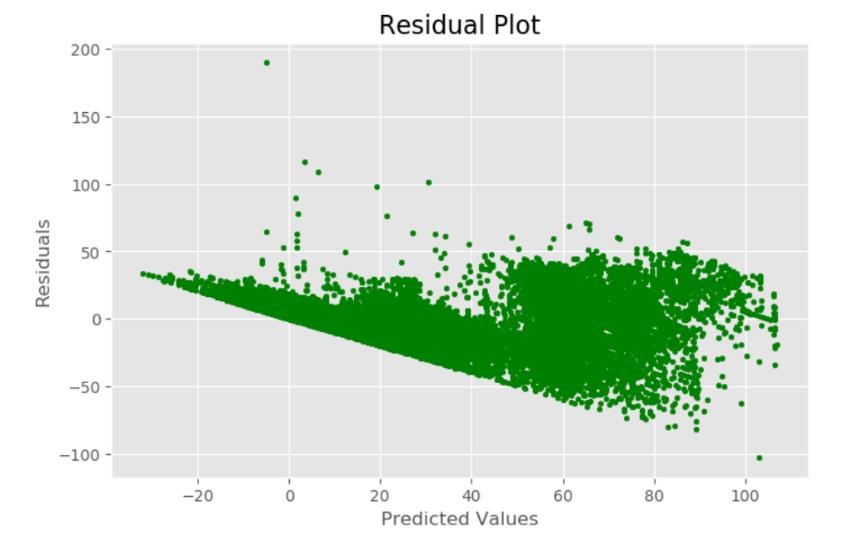


Figure 9: The above scatterplot shows a visual display of the regression lines for model 1. The yellow line represents the observed 𝑇𝑐 values plotted against each other. The blue line represents the predicted 𝑇𝑐 values for superconductors, based on the regression model. The blue dots represent predicted 𝑇𝑐 values for superconductors with oxygen plotted against their corresponding observed 𝑇𝑐 values (Python).

Figures 10 and 11:



Figures 10 and 11: These plots show the residual analysis of model 1. The residual plot on the left (figure 10) shows the residuals plotted against the predicted values. There seems to be a clear violation of the constant variance assumption because the plot takes a funnel-shaped pattern. The qq-plot on the right (figure 11) shows the observed residual quantiles plotted against the theoretical quantiles for a normal distribution. The residual quantiles roughly follow the red line except for the large quantiles, indicating skewness in the residuals and deviations from normality. These observations indicate that a transformation is needed on either the response variable or the predictor variables (Python).

Figure 12:

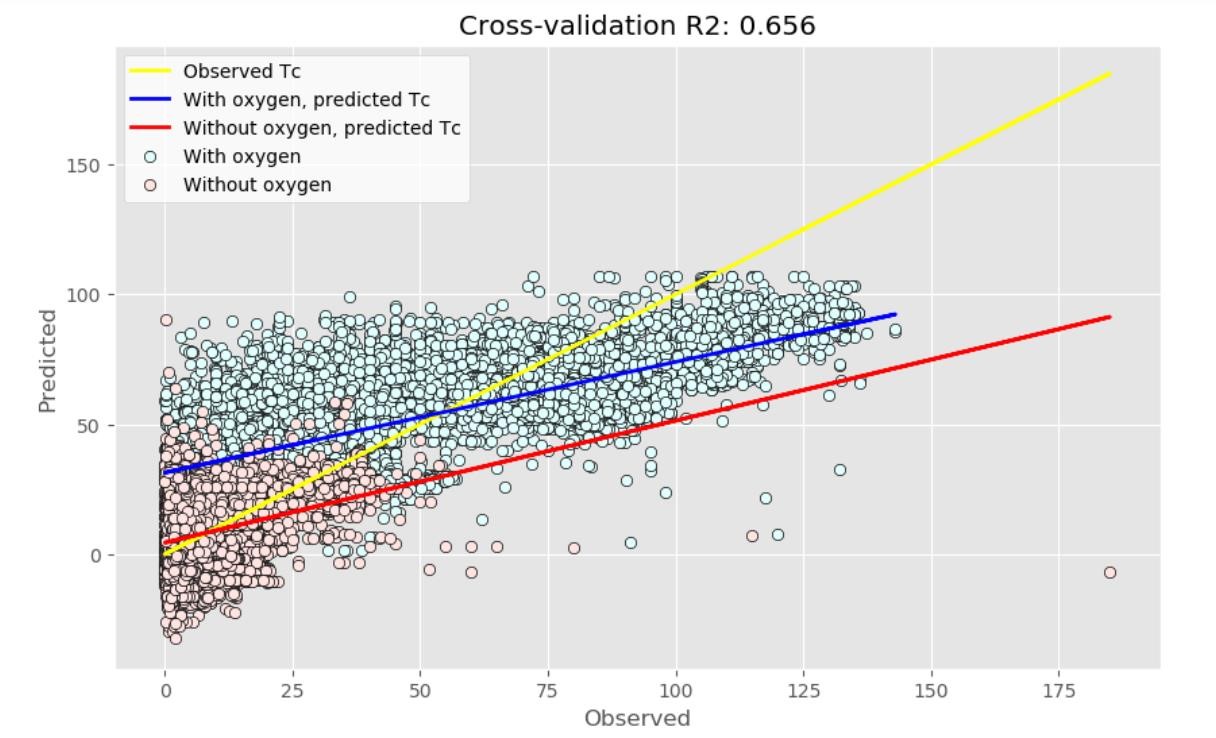
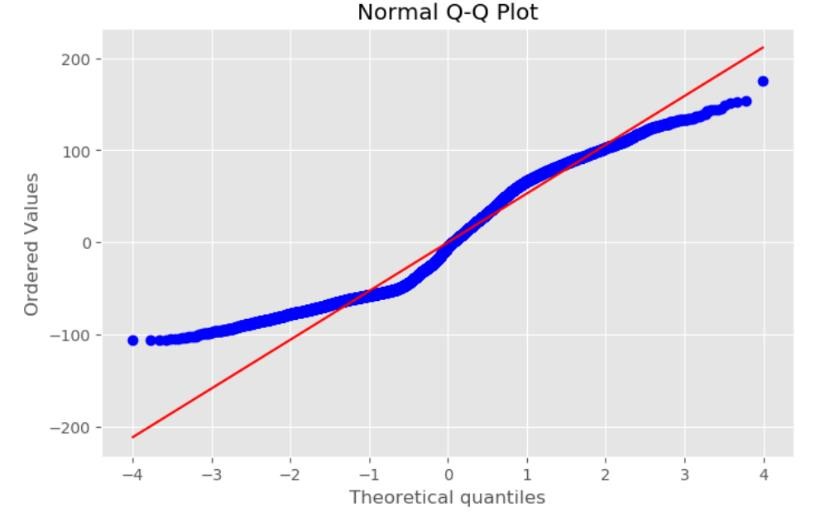
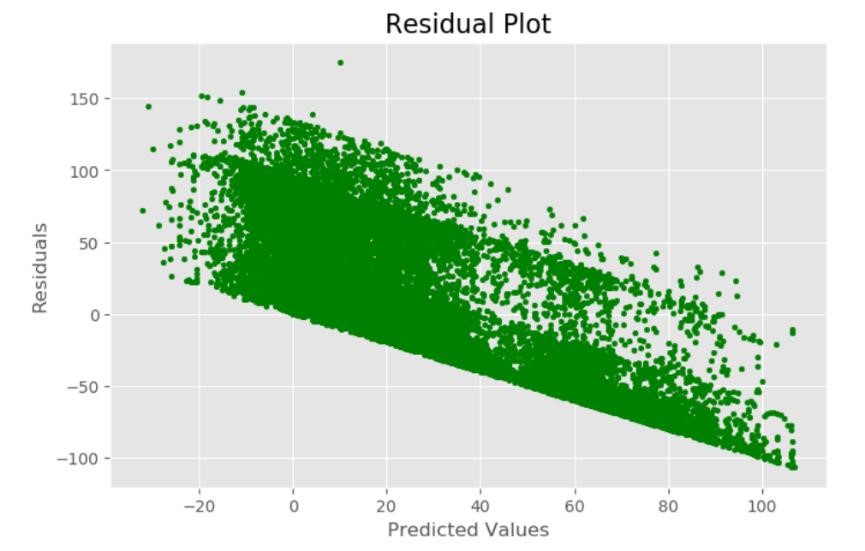


Figure 12: The above scatterplot shows a visual display of the regression lines for model 2. The yellow line represents the observed 𝑇𝑐 values plotted against each other. The blue line represents predicted 𝑇𝑐 values for superconductors with oxygen while the red line represents the predicted 𝑇𝑐 values for superconductors without oxygen. The blue dots represent predicted 𝑇𝑐 values for superconductors with oxygen while the red dots represent predicted 𝑇𝑐 values for superconductors without oxygen, both plotted against their corresponding observed 𝑇𝑐 values (Python).

Figures 13 and 14:



Figures 13 and 14: These plots show the residual analysis of model 2. The residual plot on the left (figure 13) shows the residuals plotted against the predicted values. There seems to be a clear violation of the constant variance assumption because the plot takes a funnel-shaped pattern. The qq-plot on the right (figure 14) shows the observed residual quantiles plotted against the theoretical quantiles for a normal distribution. The residual quantiles do not follow the red line, indicating the residuals are not normally distributed. These observations indicate that a transformation is needed on either the response variable or the predictor variables (Python).