|  |
| --- |
|  |
| **Final Project Report – Certified Light Bulb Analysis**  ALY6015: Intermediate Analytics   By:  Arun Radhakrishnan |
|  |

Table of Contents

[INTRODUCTION 2](#_Toc149218720)

[METHODS 3](#_Toc149218721)

[Analytical methods 3](#_Toc149218722)

[Cleaning 4](#_Toc149218723)

[DATA ANALYSIS 5](#_Toc149218724)

[EDA 5](#_Toc149218725)

[Chi-Square Test 8](#_Toc149218726)

[ANOVA 10](#_Toc149218727)

[Linear Regression 11](#_Toc149218728)

[Lasso and Ridge 14](#_Toc149218729)

[Time series against Energy Used 17](#_Toc149218730)

[Top three brands with the lowest mean energy usage 18](#_Toc149218731)

[Conclusion 19](#_Toc149218732)

# **INTRODUCTION**

The project and reports are aimed towards several learning goals to consolidate theoretical knowledge into technique and practical skills of value to analyze data using methods presented in the course. Using R to effectively process, analyze and depict data, generalized linear models, logistic regression and regularized methods for models are used to describe relationship among variables, analyze data and build a prediction model.

This paper reports the initial goals of the project, the research questions, preliminary analysis/ exploratory data analysis (EDA) of the Energy Star certified light bulbs, and the answers to questions and their methods. The report also briefly describes the variables within the data set and the key efficiency criteria to better aid domain knowledge.

The goal of the project is to understand the variations in the key efficiency criteria of bulbs relative to the technologies used in building them (product types with their use cases), extra features included, brand types, and markets which they are found. Also make predictions on some dependent variables like life hours of bulb and brightness lumens based on their respective power relative variables/specifications.

The following research questions were highlighted:

* What are the expected life hours of a bulb?
* Is there any variation in brightness among brand types and bulb types?
* Is there any association between color quality and power factor?
* Is there any association between connected lamps and three-way lamps?
* Is there any association between dimmability and being rated for enclosed features?
* Can I estimate brightness from other power related variables?
* Using lasso and ridge, which better fits the model and reduces overfitting?
* Is there any difference in power consumption for bulbs sold in US and Canada and both?
* What are the best brands of the bulbs?

The dataset used in this research is the Energy Star Certified light bulbs data set, which was sourced from data.gov and was provided by the United States Environmental Protection Agency (<https://catalog.data.gov/dataset/energy-star-certified-light-bulbs-version-2-0>). The data was collated in 2020 and updated in 2023 reporting the key efficiency criteria of energy efficient bulbs. The dataset consists of the product specifications of several bulbs, which was produced using different technologies and for different uses such as decorations or just illumination. It also highlights the markets for which the bulbs are found which is either Only U.S, only Canada or both.

# **METHODS**

## **Analytical methods**

Plots such as bar plot, histogram and scatter plots were used to get a better understanding of the distribution of the data and the variables at the EDA stage. The correlation and the correlation plot were used to understand the linearity of the numerical variables with each other.

To better understand the relation between two variables, and test if there are dependencies between such variables, the chi-square independence test was conducted on such variables, and from their resulting p-values, we understand if there are any such dependency.

ANOVA test was used afterwards to check the difference between and among groups in the data set such as brands or bulbs only sold in Canada and USA. There mean differences will also help to understand the standards for energy saving for both countries. ANOVA test will also be conducted to test the variance in the observations based on the technologies used to produce the bulbs and additional features the bulbs have. The combined effect of the technology group and the feature like being dimmable or other connection types (Bluetooth, wi-fi, Zigbee) on the power consumption in Watt will also be tested with ANOVA.

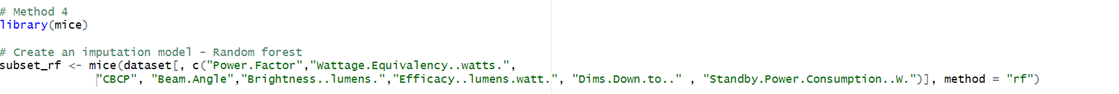
Linear regression models will be built on dependent variables such as power consumption, estimated life hours. Using lasso or ridge regression models, we can also better fit the model into a dependent variable, to better make predictions, and get a quick overview of variables that do not contribute a lot in the prediction and therefore making decisions of leave them or get rid of them.

In summary to answer the questions, chi-square test of independence will be used to determine is a variable is independent of another, and the ANOVA will be used to test variability of data among groups. Multiple models will be used with different dependent variables to predicts certain dependent variables. And using the Lasso and ridge, all will be tested with the dependent variable to understand the relationship between the dependent variable and the prediction.

## **Cleaning**

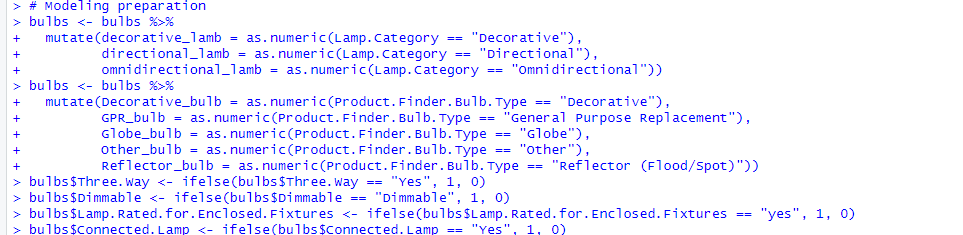
The dataset consists of 11919 observations within 39 variables which are such as brand name, model name and number, bulb type, bulb type category, lab category, technology used, energy used in watts, estimated life hours, color quality, power factor, markets, efficacy, center beam candle power (CBCP), and other features variables.

The data consisted of 29032 overall null values across all variables and rows. After a quick observation of the variables and the data, we decided to not remove any NA/null values but rather fill the nulls with substitutes. There were several ways we could replace the nulls which were using mean, mode, k-nearest neighbor, hot-deck imputation, and prediction using regression and others. We then decided to use an imputation model using random forest to replace the nulls.



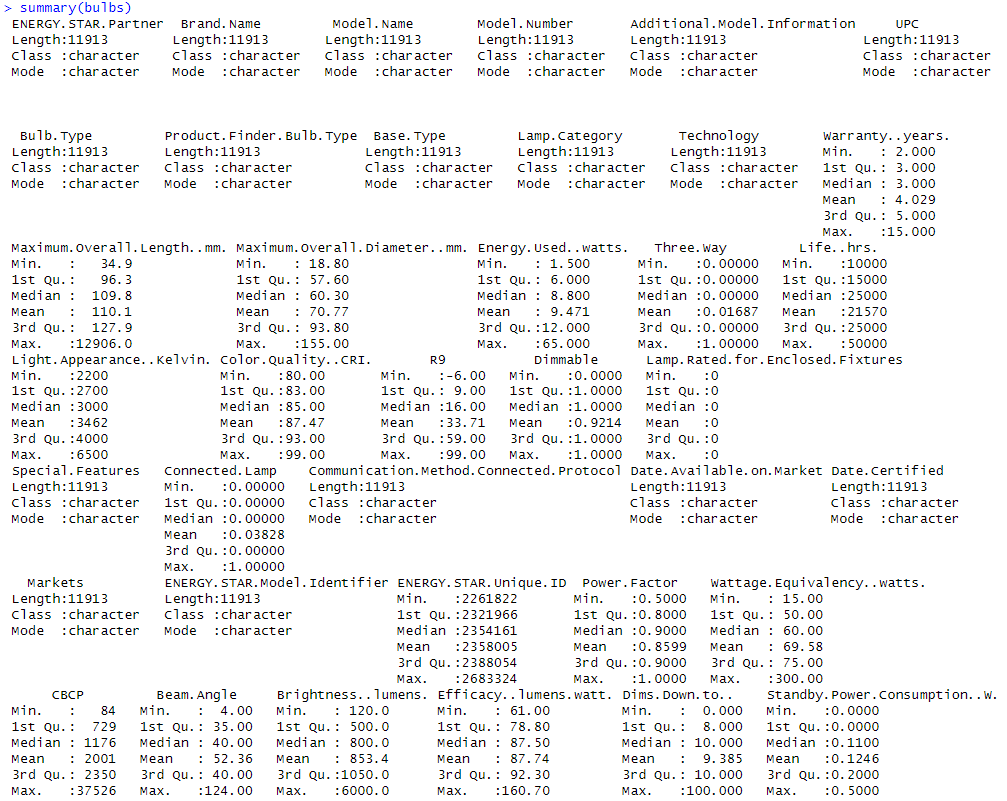
After the data was generated in a data frame, the older columns were merged with the new generated data, replacing the null as predicted.

The data contained various categorical variables with more than 2 entry types, and this will be further cleaned and manipulated to fit into the model that is to be built on the dataset. Some of the columns such as model name and brand name are currently not of concern in the EDA but could be of further reference to understanding a data observation. Weird observations in columns containing data such as ratio (0-1) and percentage were observed then removed from the dataset. Also, variables where necessary was prepared for modeling after the EDA.

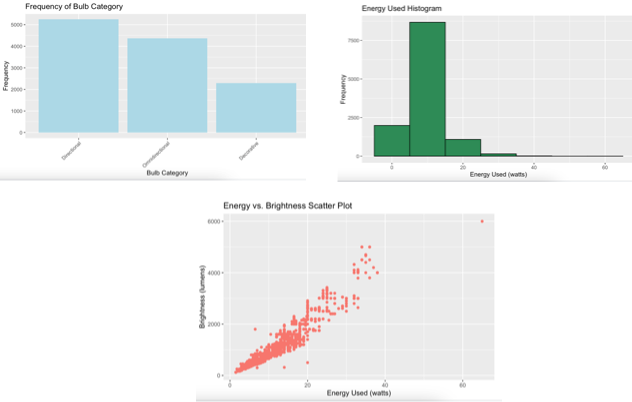


# **DATA ANALYSIS**

## **EDA**



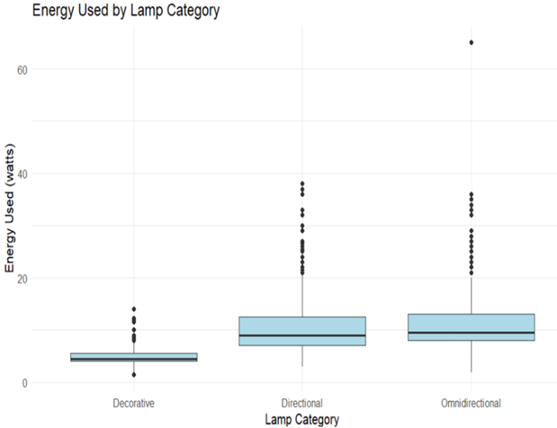
From the summary of the dataset, the mean warranty years on the bulbs is approximately 4 years. From a further observation, more than half of the bulbs had warranty years of less than 3 years. Glancing quickly at the energy used in watts, it is observed that about 75 percent of the bulbs had less than 12 Watt in energy usage, which is considerably low and of good energy saving. Also, the mean life hours of the bulbs were estimated to be 21570, which when compared to hours in a year which is 8760, the mean life in year is 2 and a half years. This could mean the warranty offered on the bulbs exceeds there expected life hours and could result in a significant loss to the company. Back to the topic on power consumption, which is the major focus of the dataset, the mean standby power consumption for the bulbs is 0.1246 in Watt. About 75 percent of the bulbs from the observation has less than 0.2-watt standby power consumption.



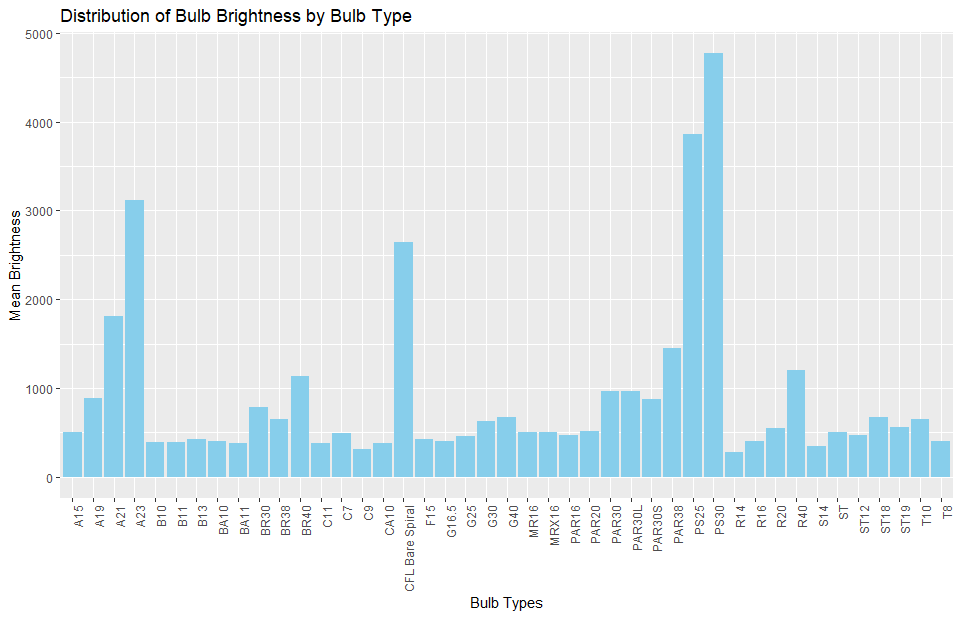
The bar plot above shows the frequency distribution of the different light categories. There are three which are decorative, directional, and omnidirectional. It helps to explore the categories of light bulbs in the dataset, for which using other tests like ANOVA we can explore the difference in means for the categories relations to variables such as brightness and CBCP. There is no specific pattern to the distribution as it is not a normal distribution, but the skewness can be evaluated.

The histogram shows the frequency distribution of the energy usage of the bulbs. With the first look, it is observed that the distribution is not normal and is skewed to the right (positive skew). This means that for our entire data, most bulbs use lower energy, with the majority less than 20-unit measurement, and the means of the distribution is greater than mode. The skewness of the distribution might also be because of outliers, but such are better examined further than removed.

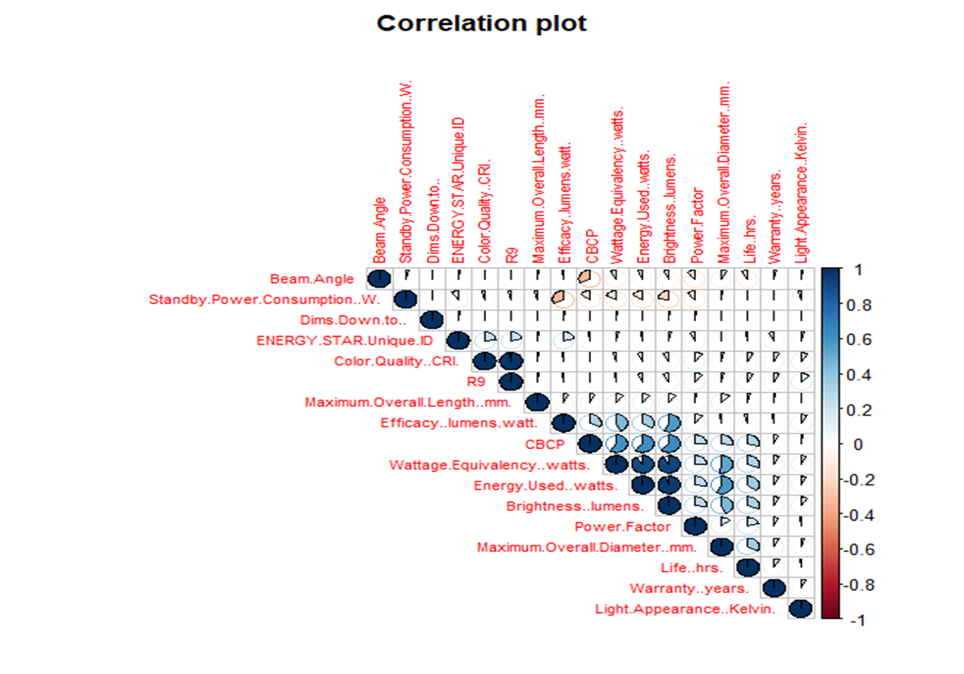
The Scatter plot shows the distribution of energy usage with the brightness produced by the bulbs. From knowledge of expertise, an increase in brightness will almost always result in an increase in power consumption. The unlikely cases would be those where more advanced technologies are used and there is lower power consumption. The scatter plot also shows an increasing linear relationship between both variables.



The energy used by lamb category box plot below gives a quick overview of distribution and statistics of the lamb category and their respective power consumption. There seems to be several outliers in the statistics, but these values are kept because further investigation into why the values are as such needs to be made. The outliers detected might be because of a different technology used in the lamps which causes it to have higher power consumption. This might also be due to the uses of the lamp, as can be seen in the plot, directional and omnidirectional lamps seem to have higher number of outliers. The outliers will be better detected and removed using the cook’s distance. For a closer insight into the energy used in watts, and the detected outliers, a distribution of the brightness of the bulb types is shown below.



From the distribution plot of the bulb types and the mean brightness shown above, it is observed that a few bulb technology types have their brightness above the range at which others are found. With a closer look into the observation data of the previously detected outlier on the energy box plot, most of the observations detected as outliers use the bulb technology of the bulbs with higher mean brightness displayed above. As will be observed in the upcoming correlation map and as seen in the scatter plot of energy used and brightness above, there is a high positive linearity between both variables. For the above stated reasons, the outliers displayed in the energy boxplot were not removed. The bar plot of the bulb types (technologies) with the energy usage also had a similar distribution to that displayed above.



The plot displays correlations between various variables in a dataset, with positive and negative correlations color-coded in a legend. Notable observations include the negative correlation between standby power consumption and most other variables, implying that an increase in standby power leads to a decrease in other variables. However, the exact cause of this relationship remains uncertain.

Variables like "dims" have minimal linear relationships with all variables except standby power consumption. Additionally, there is a negative correlation between the center beam candle power (CBCP) and the bulb's beam angle, suggesting that increasing the beam angle may lower the light intensity at the center of the bulb. This assumption aligns with optical physics principles, but further research is needed for a conclusive analysis.

## **Chi-Square Test**

Chi-square tests of independence were conducted on various variable pairs, where yielding very low p-values rejects the null hypothesis of no association between variables and failing to reject otherwise. Using a significance level of 0.05 for all tests.

**Question**: is there any association between color quality and power factor?

Using the chi-square test of independence on the two variables with the following code.

A close-up of a test

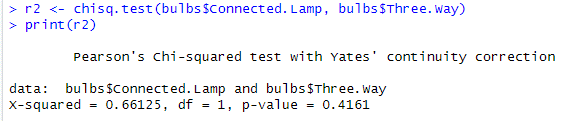
Description automatically generated

Given the null hypothesis is that there no association between both variables, and the alternative hypothesis is otherwise, from the p-value of the test above, we rejected the null that there is no association between both variables.

This helps in concluding that the color quality which is different for the several different bulbs is related to the power factor of the bulb which can be defined as the blind current being generated by the bulb. To avoid a higher power factor for the bulb, the color quality will be a factor of great consideration.

**Question:** Is there any association between connected lamps and three-way lamps?

Using the chi-square test of independence in R results in the following.

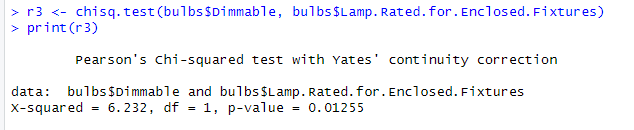


Provided the null hypothesis stating that there is no association between connected lambs and three-way lamp types, and the alternative hypothesis states there is an association between them, from the above result, we fail to reject the null. The p-value for the test is given to be 0.4161 which is way higher than the significance level of 0.05, we fail to reject the null and conclude that there is no association between both variables.

In conclusion, to produce a three-way lamp of certified energy usage criteria by Energy star, including a connection feature in the lamps will help meet the specified criteria.

**Question:** Is there any association between dimmability and being rated for enclosed features?

Also, using the chi-square independence test and the given value of alpha = 0.05, the test was conducted. The null hypothesis states there is no association between both variables and the alternative hypothesis states otherwise.



The p-value from the test is given to be 0.01255 with a degree of freedom of 1 and alpha 0.05. From the results shown above, the null that stated that there is no association between the dimmability of a lamp and the lamp being rated for enclosed features was rejected. It was rejected because the p-value of the test from our sample is lower than the level of significance and we therefore concluded there was enough evidence to reject the null hypothesis.

This verifies our assumption that to produce a lamp which is certified for enclosed features by Energy star, the bulb of the lamb must be dimmable.

## **ANOVA**

ANOVA test of variability was conducted given level of significance (alpha) to be 0.05.

**Question:** Is there any variation in brightness among brand types and bulb types?

Using a ANOVA test on the variable brand type as seen below

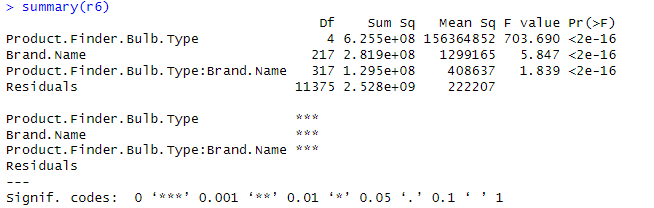
A screenshot of a computer code

Description automatically generated

The null hypothesis states there are no significant differences in brightness (lumens) among the different brands of the bulbs and alternative hypothesis states there is significant differences in brightness (lumens) among the different brands of the bulbs.

From the result shown above, In ANOVA test, the p-value is much lesser than the value of alpha. Hence the null hypothesis was rejected. The analysis shows that the brightness (lumens) of several brands of bulbs varies significantly from one another.

Rather just examining the variability of the brightness of the lamps produced by different brand partnering with Energy star, we also want to understand if there is any difference in mean of the brightness in of the bulbs produced by the brands. We then conducted a two-way ANOVA test to understand the difference among bulb groups and the interaction between both variables on the brightness.



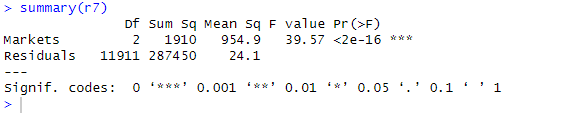
For the variability among the bulb types, the p-value for the test was given to be 2e-16 which is very low compared to our level of significance. Given the null hypothesis for the case of the bulb types being there is no difference in the means of the bulb type groups and the alternative hypothesis stating otherwise. From the test above, the null hypothesis that there is no difference in variation was rejected, concluding that there is a difference in mean for the bulb types as well.

Finally, to understand the interaction of both variables on the brightness, the null hypothesis states there is no interaction between the brand type and bulb type on the brightness of the bulb, and the alternative hypothesis states there is an interaction between both variables. Also, from the test above, the p-value of the test is very low, and the null hypothesis was therefore rejected. This helps in concluding that, there is a difference in the brightness of the bulbs produced by the brands partnered with Energy star. So, to meet the energy saving certification criteria, one does not need to follow a general rule of brightness or bulb criteria to produce.

**Question:** Is there any difference in power consumption for bulbs sold in US and Canada and both?

The null hypothesis states that there is no difference between the energy consumption in watts for bulbs sold in the different market groups.

The alternative hypothesis states that there is a significant difference between the energy consumption in watts for bulbs sold in USA and Canada.

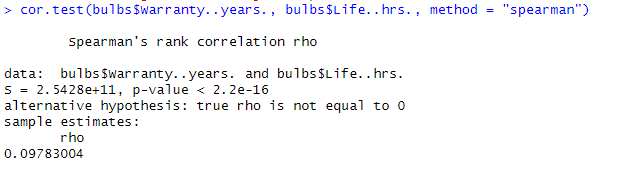


The p-value of the test is very low compared to the level of significance, therefore, the null hypothesis which states that there is no difference in means of energy consumption for the market groups was rejected. This helps to conclude that there is a difference in the energy usage of the bulbs found in the market groups. There in a difference in the means of energy usage of bulbs certified energy saving sound in the USA and Canada. The variability might be random or due to policies unknown in both markets, but with more data, further investigation can be done to ensure requirements are met to enter either market.

**Spearman Correlation**

The spearman’s correlation was used because both variables are non-normally distributed. It helps to measures the degree to which a change in one variable is associated with a monotonic change in the other variable. With the value of Rho (spearman correlation constant) ranging from -1 to 1 (negative and positive correlation respectively), a test was conducted to examine the warranty years and life hours of the bulb. This analysis resulted from the observation in the summary which was initially noticed that the mean warranty years is more than the mean life hours of the bulbs.

The null hypothesis states that there is no significant correlation between the warranty years and the life hours of the lamp. The alternative states that there is a statistically significant correlation between both variables.

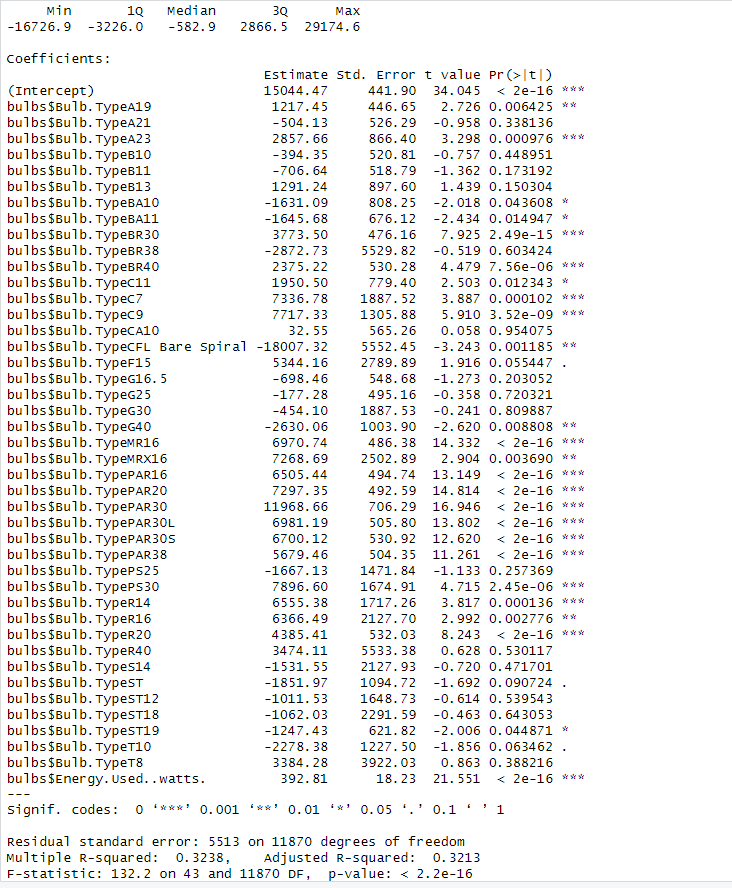


The p-value of the test is very low, therefore rejecting the null hypothesis that there is no significant correlation between both variables. There was enough evidence to reject the null but that is not the only important observation. The value of Rho is given to be approximately 0.09783 which means there is a very weak positive correlation between life hours and warranty years. This could result in losses for the company if the warranty years for the bulb is less than the expected life hours of the bulb.

## **Linear Regression**

**A computer code with text

Description automatically generated with medium confidenceQuestion:** What are the expected life hours of a bulb?

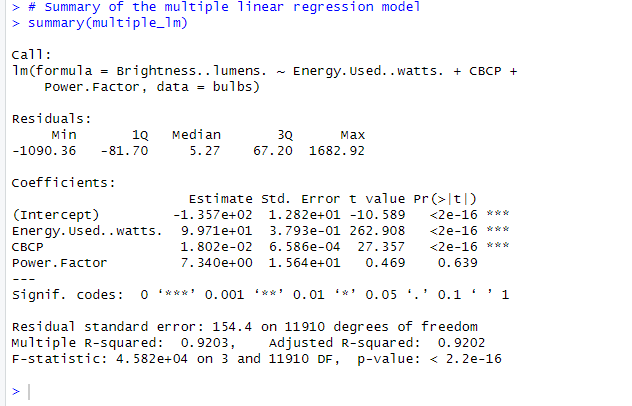


A screenshot of a graph

Description automatically generated

The linear regression model has great overall statistical significance, with a very low p-value for the F-statistic, showing its capacity to explain a significant percentage of the variation in the log-transformed 'Life..hrs.' Some predictor factors stand up as extremely important in predicting 'Life..hrs.', such as 'Energy.Used..watts.' and various 'Bulb.Type' categories. The R-squared value of 0.3431 for the model indicates that it accounts for about 34.31% of the variance, indicating a decent match. It is important to remember, however, that the model may not represent all of the variance in the dependent variable. It is vital to carefully read the coefficients of significant factors in order to understand their unique influence on 'Life..hrs.'

Question: Can I estimate brightness from other power related variables?



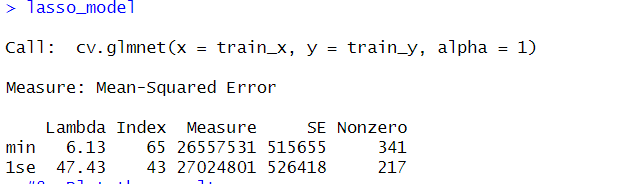
The model is extremely significant, and the output gives the coefficients for each independent variable, enabling you to assess their influence on the dependent variable "Brightness..lumens." "Energy.Used..watts." and "CBCP" show a high positive association with brightness in this scenario, however "Power.Factor" does not appear to be a relevant predictor in this model.

## **Lasso and Ridge**

Question: When predicting the lifespan of LED lightbulbs, can Lasso and Ridge regression techniques help improve model accuracy and reduce the impact of irrelevant features?

Null Hypothesis: Lasso and Ridge regression can improve lifespan prediction models and alternative hypothesis states otherwise.

Lasso Regression



Inferences:

• The y-axis denotes the mean squared error (MSE), while the x-axis portrays the logarithmically transformed lambda values. On the upper section of the graph, non-zero coefficients are depicted.

• The red dots, thoughtfully positioned throughout the graph, represent the error estimates. The lines extending above and below these red dots delineate the confidence intervals of the error matrix. These red dots are derived from the Lasso method's cross-validation.

• Within the plot, two vertical lines come into view. The first line signifies lambda.min, which is positioned on the left side, while the second line designates lambda.1se, found on the right side.

• the plot reveals that lambda.min appears to fall within the range of 6.13, whereas lambda.1se is situated at a value close to 47.43.

A graph with numbers and a line

Description automatically generated

Ridge Regression

A white background with black text

Description automatically generated

Inferences:

• Minimal disparity is observed in the coefficients between lambda.min and lambda.1se.

• The lambda.min is apparent at approximately 236, while the lambda.1se is evident at around 2416.

• The constant value of 354 at the top signifies the number of variables that remain unaltered across both lambda.min and lambda.1se.

A graph with numbers and a line

Description automatically generated

Comparing the performance of Lasso and Ridge from the R² Value and the MSE

R-squared (R²) Value:

Interpretation: R² measures the proportion of the variance in the dependent variable (target) that is explained by the independent variables (features) in the model. It ranges from 0 to 1.

A higher R² value indicates that a larger proportion of the variance in the target variable is explained by the model. It suggests a better fit of the model to the data.

Lasso has an R² value of 0.7144048 , which means it is a good model that explains most of the variance.

Mean Squared Error (MSE):

Interpretation: MSE measures the average squared difference between the predicted values and the actual values (i.e., the residuals). It quantifies the model's error in predicting the target. A lower MSE indicates that the model's predictions are closer to the actual values.

Lasso’s MSE is 228159773 indicates a better fit of the model to the data.

**Based on the above results, we can accept the null hypothesis that lasso can improve the performance of the model in predicting bulb lifespan.**

## **Time series against Energy Used**

A graph with blue lines and text

Description automatically generated

The linear trend line is relatively flat, it suggests that there's no significant change in energy usage over time. Light bulbs remain consistent in their energy efficiency.

**Time series against life of bulbs**

A graph with numbers and lines

Description automatically generated

The linear trend line (in red) slopes downwards, it suggests that life of the bulb has been decreasing over time for the certified light bulbs. This could indicate that newer bulbs have less life.

## **Top three brands with the lowest mean energy usage**

A blue rectangle with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

* When choosing on the best brand among the top five with the lowest mean "Energy.Used..watts." values, it's critical to examine things other than energy usage. "ELITCO" is the most energy-efficient choice in terms of "Energy.Used..watts." However, "Archipelago" and "ENVISION LED" are also worth considering because to their low energy consumption and possible benefits in terms of brightness and power factor. The decision should be guided by individual requirements and priorities. If energy efficiency is the most important criterion, "ELITCO" is the obvious option; but, if we consider value a balance of energy consumption, brightness, and power factor, "Archipelago" and "ENVISION LED" are good rivals. Finally, what constitutes a "good" brand is determined by individual criteria and the things that are most important.

# **Conclusion**

We conducted a detailed investigation of Energy Star certified light bulbs in this research, employing a variety of statistical approaches and data processing techniques. The study questions looked at things like bulb efficiency criteria, brand and product kinds, market performance, and their connections. Our data cleansing and exploratory data analysis (EDA) revealed crucial properties of these bulbs, such as warranty years and energy usage, as well as brightness and power consumption. To discover connections and variances between variables, we used chi-square and ANOVA testing. To measure bulb life hours and forecast brightness, linear regression models were utilised. Model accuracy was enhanced using the Lasso and Ridge regression approaches, notably in forecasting bulb lifespan. Time series analysis identified patterns in energy use and bulb life, revealing useful information. This project demonstrates the capability of data analysis using R to acquire practical insights from real-world data, assisting in informed lighting sector decision-making.

**REFERENCE**

GeeksforGeeks. (2023). What is Exploratory Data Analysis. *GeeksforGeeks*. <https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/>

Facer, C. (2022, August 23). *How to create a correlation matrix in R - Displayr*. Displayr. <https://www.displayr.com/how-to-create-a-correlation-matrix-in-r/#:~:text=A%20correlation%20matrix%20is%20a,negative%20correlations>).

Porras, E. M. (2022b). *Linear regression in R tutorial*. <https://www.datacamp.com/tutorial/linear-regression-R>

**APPENDIX**

