D-214

Task-3

Presentation of Findings

22nd February 2022

Presented By : Harshal Parikh
Student ID: 007091749
Program : Masters in Data Analytics
Program Mentor : Jared Knepper
Contact : *hparikh@wgu.edu*

`

Executive Summary and Implications

Statement of the Problem

We have a total of 8 different components affecting the Heating and Cooling load of the building. These parameters led to draft the statement of the problem.

“Could we identify the principal components that have the greatest impact on the Building Energy Efficiency”

Hypothesis

The orientation of a building statistically significantly impacts the Heating and cooling of residential buildings.

We anticipate that we will be able to create a statistical model with a moderate to high success within a certain confidence interval. The model will predict the most important principal components of interest relevant to the Heating and Cooling Load of the building. The model will also provide coefficients (Weights) for all the features in the regression model. We intend to identify the component / components which have a large coefficient value.

Summary of the Data Analysis Process

Prior to performing Data Analysis, it is crucial to ensure the following steps are performed.

- Ensuring the libraries and packages have been imported and the csv file has been read
- Renaming the columns and performing steps to ensure the data has been cleaned.
- Cleaning the data would constitute: Checking for empty rows/columns and Duplicates

The variable used for PCA, and regression analysis was Orientation. Summarizing the data analysis process is a multistep process.

- Defining the variables followed by standardizing the data
- Creating a covariance matrix and performing an eigen decomposition on the covariance matrix.
- Performing an PCA. Listing the descending sorted eigenvalues, fitting a standardized matrix of features, and printing the explained variance ratio.
- Performing an OLS regression and developing an Initial Estimated Regression Equation

Outline of the Findings

The Building Energy Efficiency Dataset has a total of 768 rows and 10 columns. This dataset has 8 attributes which help determine the Heating and Cooling load of the building.

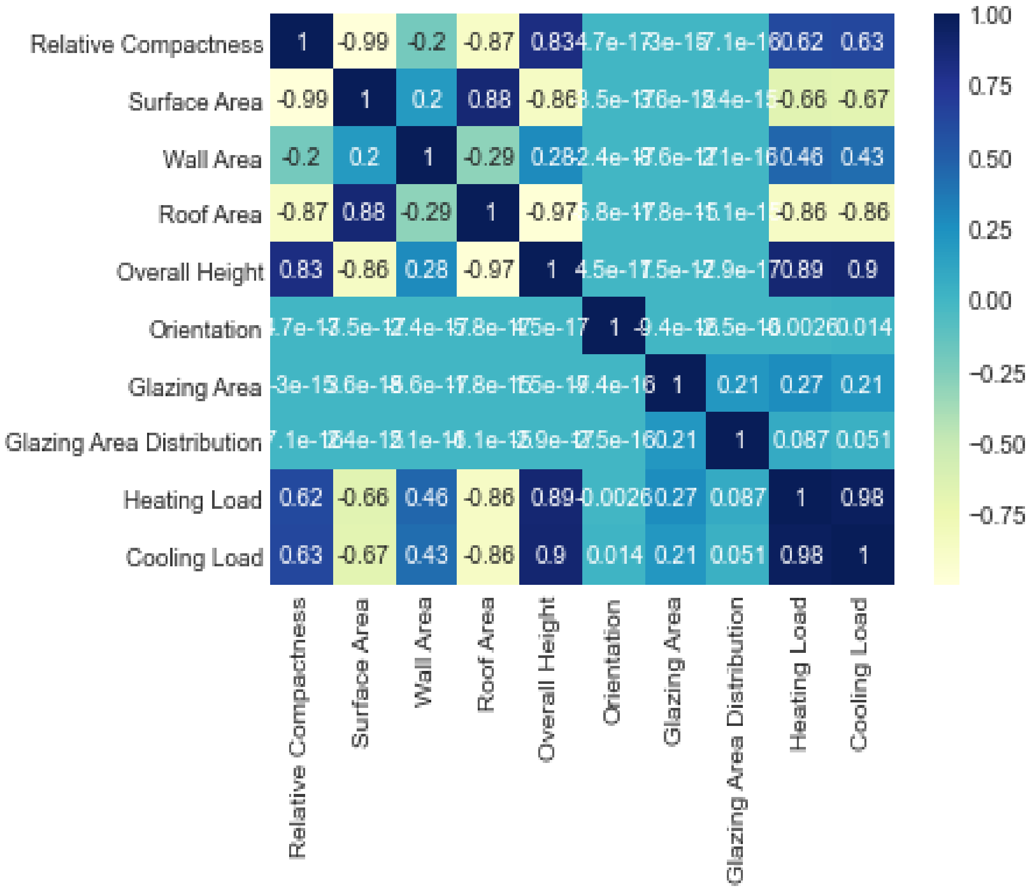
On visualizing the Scree Plot we get a total of 5 components which indicate 100% of the variance. The OLS regression gave us R-squared (uncentered) value of 0.908 and an adjusted R-squared value of 0.907. We got an Alkaline Information Criteria (AIC) of 2362 and an Bayesian information criteria of 2399. The variance vs component plot indicate that a total of 3 components have a variance of 1.0. Since we get the smallest eigenvalue of 2.27 e-25, we get an indication of strong multicollinearity problem or the design matrix being singular.

The initial multiple regression model gives **Y** = 1.8766***Relative Compactness** + 0.0022***Surface Area** + 0.0004***Wall Area** + 0.0009***Roof Area** + 0.0131***Overall Height** + 0.1748***Glazing Area** + 0.0071***Glazing Area Distribution** -0.0438* **Heating Load** + 0.0476***Cooling Load**.

Summarizing the findings we get an understanding that the first 5 components explain 100% of the variance.

The seaborn heatmap gives the following correlation findings with respect to the heating load:-

- Heating Load and Relative Compactness is 0.62
- Heating Load and Wall Area is 0.46
- Heating Load and Overall Height is 0.89
- Heating Load and Orientation is 0.0026



This gives a clear indication that the orientation of the building does not statistically significantly affect the heating load of the building.

With respect to the Cooling Load we observe the following correlation findings: -

- Cooling load and Relative Compactness is 0.63
- Cooling load and Wall Area is 0.43
- Cooling load and Overall Height is 0.9
- Cooling load and Orientation is 0.014

This also gives an indication that the orientation of the building does not statistically significantly affect the cooling load of the building.

We also observe that the Overall height has a strong correlation of 0.89 and 0.9 with the heating and cooling load of the building which is a direct indicator of the building energy efficiency.

Explanation of the limitation and techniques and tools used

The tools used to gather the dataset was the UCI Machine Learning Repository. The limitation of obtaining a public dataset is it might be out of date and not relevant in a current scenario. Another challenging limitation is the possibility of a Mosaic Effect. Mosaic effects is an indication of gathering personally identifiable information from multiple open sources which affects an individual.

The technique used to analyze the data was Principal Component Analysis (PCA). The limitations of this technique was that the independent variables become less interpretable, data standardization is must before PCA and Information Loss

Summary of Proposed Actions

Additional Actions that could be performed along with principal component analysis is

- Obtain more relevant and up-to-date information along with conducting Customer surveys and looking for ways to obtain additional data to improve stakeholder insight.

- Perform an analysis to see if additional parameters like glazing area, glazing area distribution, Wall Area, Roof Area and Surface Area prove to be statistically significant.

- Since time series analysis provides a specific way of analyzing the sequence of data points which were gathered over an interval of time that would be an additional proposed action.

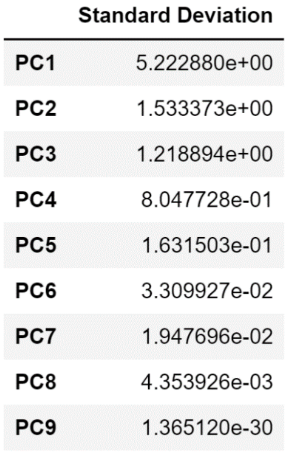
Expected Benefits of the study [Specific and Quantitative]

Principal Component Analysis has been widely used to reduce the dimensionality of the data. We were able to identify

- 2 components give us more than 70% of the variance.
- 4 Components were able to give 95% of the variance
- 5 components give 100% of the variance of the data.

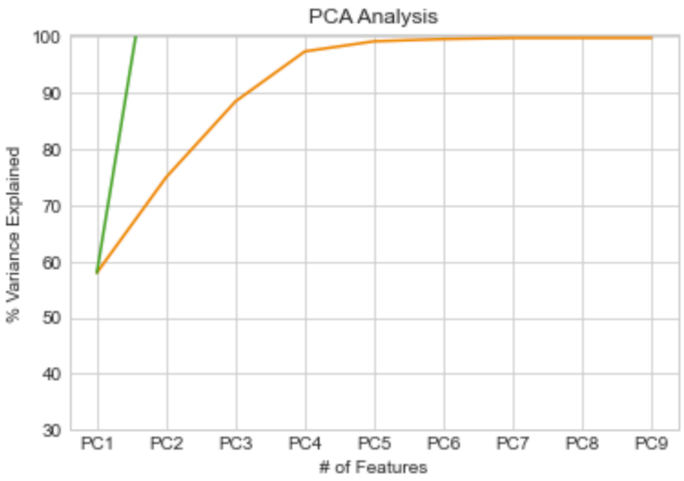
Heatmaps provide an easy-to-use interface to visualize the data.

Observing PCA along with heatmaps, I was able to identify that our assumption of the hypothesis that orientation had a major impact on cooling and heating load of the building was not plausible. On the other hand, we also got an understanding that the overall height correlates a 0.89 and 0.9 to the heating and cooling load of the building.



	Standard Deviation
PC1	5.222880e+00
PC2	1.533373e+00
PC3	1.218894e+00
PC4	8.047728e-01
PC5	1.631503e-01
PC6	3.309927e-02
PC7	1.947696e-02
PC8	4.353926e-03
PC9	1.365120e-30

The study would also give stakeholders an understanding on what component has the highest impact in improving the energy efficiency of the building



Panopto Recording

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=387afaea-09bd-4961-83f5-ae440177c2b7

References

1. What are Pros and Cons of PCA. Retrieved from https://www.i2tutorials.com/what-are-the-pros-and-cons-of-the-pca

2. Time Series Analysis: Definitions, Types, Techniques and when it is used. Retrieved from https://www.tableau.com/learn/articles/time-series-analysis#:~:text=Time%20series%20analysis%20is%20a,data%20points%20intermittently%20or%20randomly