­

**Features Impacting Residential Solar Project Costs in San Diego, California**

Christopher Fischer - 011933891

College of Information Technology, Western Governors University

Dr. Daniel Smith

November 23, 2024

Features Impacting Residential Solar Project Costs in San Diego, California

# Research Question

San Diego, California, boasts one of the highest solar adoption rates in the United States. However, to reach the city’s goals of net-zero emission of greenhouse gases by 2035, the rate of adoption will need to accelerate. (City of San Diego, 2024) One of the most consistently reported barriers hindering residents from choosing to go solar is the cost of installing a system. (National Renewable Energy Laboratory, 2016) Furthermore, developing an installation cost estimate for a home is time-consuming, requiring an estimator with significant expertise to spend several hours to several days gathering information and calculating costs. (Plugged In Academy, 2023) If a reliable cost-prediction model can be built, a customer-facing tool could be offered to allow a homeowner to quickly receive a rough estimate after answering just a few questions. Such a tool could help lower reluctance to begin the process (Tong, 2016) and help San Diego reach its clean energy goals. The first step in producing such a model is to identify features of a solar project that have a statistically meaningful impact on the target variable.

With this context in view, this study attempts to answer the question: *To what extent can the features of a proposed residential solar project (system size, single/multi-phase, self-installed, number of modules, module capacity, module efficiency, and battery capacity) affect its total installed cost?* The study’s hypotheses are:

**H0** – The features of system size, single/multi-phase, self-installed, number of modules, module capacity, module efficiency, and battery capacity do not statistically impact the total project cost.

**Ha** – The features of system size, single/multi-phase, self-installed, number of modules, module capacity, module efficiency, and battery capacity statistically impact the total project cost.

# Data Collection and Preparation

The data used for this study come from the “Tracking the Sun” project run by the Lawrence Berkeley National Laboratory. The dataset published on the project’s website (<https://bit.ly/trackingthesun-2024>) contains data on over 3.7 million solar projects installed in the US through 2023. (Electricity and Markets Policy Group, 2024)

One advantage of selecting this dataset for study is its reliability. 72 utilities and state utility regulatory agencies across 31 states supplied this data to the laboratory. While some of the data is available through other publicly accessible channels, much of it was provided under special agreements allowing the laboratory to share it freely with the public. (Barbose et al., 2024)

One disadvantage of selecting this dataset for study is its size and variability. The volume of data exceeds the capability of this researcher to perform studies. Additionally, each entity that participated in the project supplied varying subsets of features in their portion of the data. This important limitation means some features may not be suitable for studies across the entire population due to their sparsity. For these reasons, the dataset will be reduced to include only residential solar projects installed in the greater San Diego, California region during 2023. The variability of feature availability is avoided by delimiting the data to include only one region. Further, since the target variable of this study is price, restricting data to just the most recent year is sensible due to the dual factors of marked inflation in the US over the last several years and the ever-downward trend of prices for electronic components. This dataset contains the city name and zip code for the location where each project is installed. Since these attributes will not be studied, the data will be anonymized by setting all city values to “San Diego” and all zip code values to “92101”. The resulting dataset consists of 43,265 observations.

Despite reducing the dataset to the targeted observations, it was found that a substantial number of features exhibited excessive sparsity, with values available for only a tiny subset of observations. Therefore, a feature reduction process was implemented. This process involved selectively retaining only those features with sufficient density. As a result, the dimensionality of the dataset was reduced, supporting a more accurate model and reduced computational load. (Jafari, 2022, p. 384) However, it should be noted that one sparse feature, *Battery\_rated\_ capacity\_KW*, was retained in the study. This decision was made based on the domain knowledge that many solar projects do not include a battery storage component and that adding battery storage can significantly impact the cost of a project. (Ramasamy et al., 2022)

The following features were selected for inclusion in the study:

| Variable Name | Variable Type | Data Type | Description |
| --- | --- | --- | --- |
| Total\_installed\_price | Target | Continuous | The total price of the solar project |
| PV\_system\_size\_dc | Predictor | Continuous | The total rated output of the solar project |
| Multiple\_phase\_system | Predictor | Nominal | A Boolean that indicates if the system is single or multi-phase |
| Self\_installed | Predictor | Nominal | A Boolean that indicates if the system was self- or professionally installed |
| Module\_quantity\_1 | Predictor | Discrete | The number of solar panels included in the project |
| Nameplate\_capacity\_module | Predictor | Discrete | The rated output of each solar panel |
| Efficiency\_module\_1 | Predictor | Continuous | The rated efficiency of each solar panel |
| Battery\_rated\_capacity\_KW | Predictor | Continuous | The rated capacity of the battery system, if one is included in the project |

Table - Study Variables

# Data Extraction and Preparation

The data science team will develop the code for this study using Python. In its nearly 35 years, Python has become one of the most widely used general-purpose programming languages. Python is backed by a massive community of developers and a vast collection of libraries that extend the core capabilities of the language. (Datacamp, 2022) Another reason for selecting Python is that, while R is an excellent choice for interactive studies, Python is better suited for deployment on production servers as part of a data pipeline. (WGU Information Technology, n.d.) SAS was also considered and discarded because it is closed-source, proprietary, and costly. (Yadav, 2023)

The development environment for this study uses Jupyter Notebook in the Visual Studio Code editor. This pairing enhances data analysis by combining an interactive notebook experience with advanced code editing features. Analysts can run code cells, view outputs, and leverage powerful development aids like IntelliSense and version control. Notebooks may also be annotated using industry standard Markdown providing valuable commentary in line with code, results, and visualizations.

The following libraries are used to support this study:

|  |  |
| --- | --- |
| Module | Purpose in this Study |
| Pandas | Provides tools for managing datasets |
| Matplotlib.pyplot | Offers tools for visualizing data and results |
| Seaborn | Additional visualization capabilities |
| Shap | Visualizes feature contributions to the results of a model |
| Statsmodels | Provides a set of tools for creating and evaluating models, particularly well-suited for statistical analysis and model interpretation (Srinivasan, 2020) |
| Sklearn | Implements scaling and additional regression methods |

Table - Python Libraries Supporting Study

## Prepare the Environment

A screen shot of a computer

Description automatically generated

## Load and Examine the Data

A screenshot of a computer

Description automatically generated

## Evaluate Data Sparsity

A white rectangular object with black text

Description automatically generated

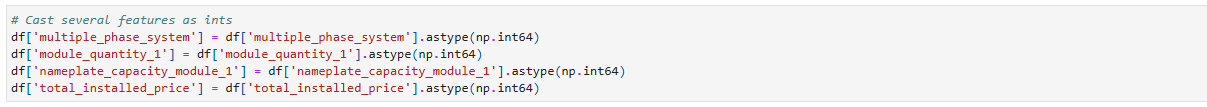
## Evaluate and Clean Missing Values

A screenshot of a computer

Description automatically generated

For the purposes of this study, *battery\_rated\_capacity\_kW* is likely to contribute meaningfully to the cost of a solar power project. Additionally, this feature’s sparsity is much greater than the remaining features. Thus, it is reasonable to conclude that projects with no battery component were indicated by leaving the value blank. Therefore, missing values are replaced with zeros so that they may accurately contribute to the regression model.

## Transform Features



Several features in the dataset were represented as floats despite having no decimal component to any values. Therefore, these features were expressed as integers for clarity.

## Identify and Treat Potential Outliers

A white background with colorful text

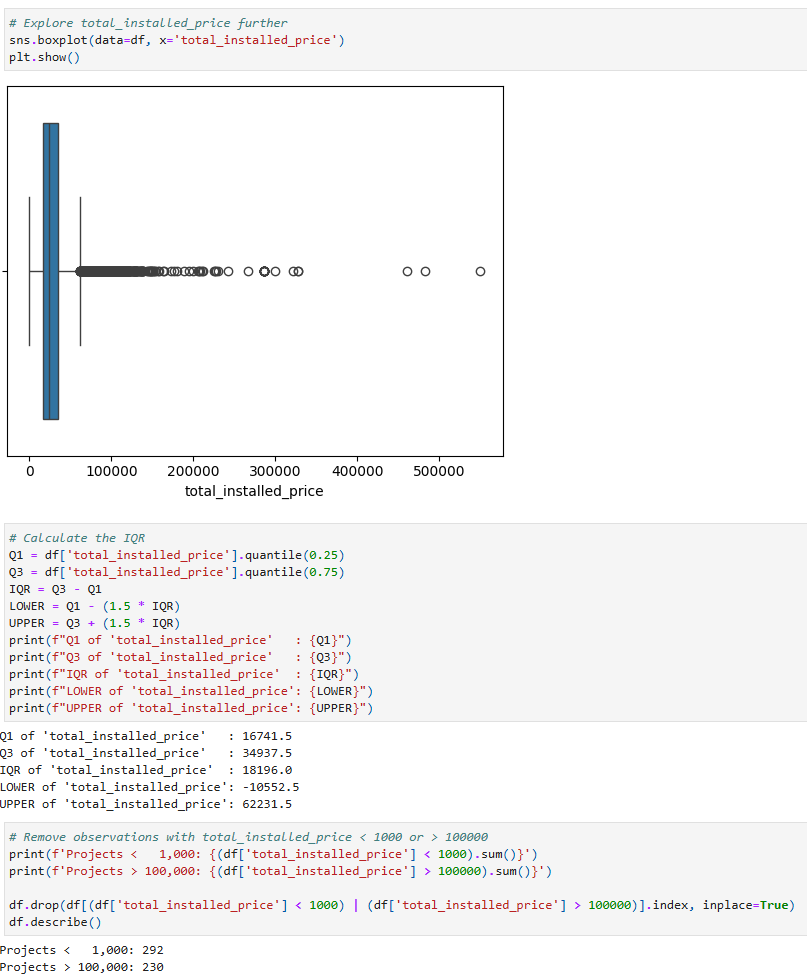
Description automatically generated with medium confidence

Potential outliers were identified by using z-scores. (Gulati, 2022) A relatively small number of outliers are noted in the independent variables.

A screenshot of a computer

Description automatically generated

Given the small number of observations with a non-zero value for *battery\_rated\_capacity\_kW*, this feature was excluded from outlier processing. All other observations with z-scores above 3 were removed, comprising less than 3% of the observations.



The target variable, *total\_installed\_price*, warranted detailed investigation to decide which observations should be excluded from the study. Calculating the IQR for the variable indicates that a range of $0 - $62,232 may be suitable. Analysts consulted reliable pricing data published by the US Department of Energy to validate this range. (Basore et al., 2024) This data suggests a range that goes upwards towards $100,000. Further, for the purposes of this study, projects with a cost of zero were considered unlikely to be accurate. Therefore, a $1,000 - $100,000 range was used to exclude an additional 522 observations from the dataset.

## Exploratory Data Analysis

A good first step in any data analysis project is to explore and visualize the study variables. This will help the team understand the data distribution and identify any relationships between variables. (Wu & Coggeshall, 2012)

A graph of a graph

Description automatically generated with medium confidence

The target variable of the study exhibits a degree of right skew. This is consistent with the findings of the boxplot visualized in the outlier remediation section.

A screenshot of a computer

Description automatically generated

Several dependent variables exhibit a similar degree of right skewness. This suggests the possibility of a correlation between these variables and the target. However, further study will be required to make a conclusive determination of correlation.

A graph of multiple phase system

Description automatically generated

Not much information can be gleaned from these frequency charts.

A screenshot of a graph

Description automatically generated

These scatter plots with regression lines more clearly show the potential for correlation between the target variable and *PV\_system\_size\_DC*, *module\_quantity\_1*, and *battery\_rated\_capacity\_kW*.

A graph of a diagram

Description automatically generated with medium confidence

The boxplots show a material reduction in the mean and IQR of *total\_installed\_price* for projects where *self\_installed* is 1. However, other variables may explain the reduction. Therefore, no conclusion can be made on this basis alone.

A screenshot of a computer

Description automatically generated

The heatmap of correlation coefficients reveals two important factors. First, a nearly perfect correlation exists between independent variables *module\_quantity\_1* and *PV\_system\_size\_DC*. Were both variables retained in the model, they would have an unwanted influence on the results due to strong multicollinearity. Therefore, *module\_quantity\_1* will be removed from the study. Second, a moderate to strong correlation between the target variable and the independent variables *PV\_system\_size\_DC* and *battery\_rated\_capacity\_kW* is exhibited.

# Analysis

The stated purpose of this study is to determine the extent to which certain characteristics of a residential solar project impact the project's cost. In the vernacular of data scientists, the study seeks to determine independent variables that contribute statistically to the model. To accomplish this outcome, a linear regression model was created. Linear regression is a simple and effective tool for estimating the degree of relationship between the predictors and the target. (Ellis, 2022) As this work focuses on explaining relationships in the existing data set rather than predicting outcomes, the *statsmodels* library was selected to implement the linear regression model over *scikit-learn*. (Srinivasan, 2020)

One advantage of using linear regression for this study is that the results can be easily interpreted and explained to varied audiences. Coefficients succinctly describe the magnitude of change in the target variable for each unit of change in the predictor variable. At the same time, p-values demonstrate the significance of each variable’s influence on the outcome. However, linear regression is not without disadvantages. One important drawback is that outlying values can easily skew linear regression models. (Ellis, 2022) Therefore, proper identification and treatment of outliers was a key component of the data preparation process for this study.

## Identify and Remove Features with Strong Multicollinearity

A screenshot of a computer program

Description automatically generated

Since multicollinearity can lead to model overfitting (Bobbit, 2019), features with a variable influence factor (VIF) greater than 5.0 were marked for removal. Since *module\_quantity\_1* was previously removed from the study, no further high-VIF features were found.

## Backward Stepwise Feature Elimination

A screenshot of a computer code

Description automatically generated

Ultimately, stakeholders desire a model that is no more complex than necessary to achieve the desired outcomes. One way to simplify a model is to remove features that are not statistically significant to the model. This significance is evaluated by examining the p-value of each feature. The p-value is “the probability of obtaining the observed results of a test, assuming that the null hypothesis is correct.” (Srinidhi, 2020) Recalling that the null hypothesis for this study states that the selected independent variables do not meaningfully impact the dependent variable, the p-value is the probability that each variable does not contribute to the cost of a solar project. The higher the p-value, the more likely the null hypothesis is correct. Correspondingly, the lower the p-value, the more likely the alternate hypothesis is correct. Therefore, the team applied backward stepwise feature elimination to remove any independent variables with a p-value greater than 0.05 from the model. (Srinidhi, 2019) Surprisingly, no predictor features were eliminated.

## Feature Scaling

A white background with black text

Description automatically generated

Features having wildly different scales can impede many machine-learning algorithms. Even those algorithms that are not impacted by such conditions can benefit from scaling features prior to model fitting. (Idriss Jairi, 2024) Therefore, all remaining features, both predictor and target, were scaled with the standard scaler.

## Create and Fit the Model

A screenshot of a computer

Description automatically generated

The coefficients table reveals that each of the remaining predictor variables contributes in a statistically significant manner to the model’s results. Further elaboration on this table will be found in the Data Summary and Implications section below.

## Test for Heteroscedasticity

A screenshot of a computer program

Description automatically generated

Homoscedasticity is the uniformity of the variance of values in a dataset. Since ordinary least squares regression assumes homoscedasticity, a Breusch-Pagan test was performed on the resulting model. (Frost, 2019) The results indicate that no heteroscedasticity was observed.

## Visualize Model Statistics

A screenshot of a computer program

Description automatically generated

A comparison of blue and red graphs

Description automatically generated

Unfortunately, the R2 suggests that only 52% of the variability in the target variable can be explained by the selected predictors. The plots of residuals and actual vs. fitted seem to support this result.

# Data Summary and Implications

This study aims to identify which features of a residential solar project contribute materially to the total installed cost. One must examine two measures for each predictor to determine which features do so. First, is the predictor statistically meaningful, and second, to what degree does a change in the predictor alter the value of the response? (Gallo, 2016) The model summary from *statsmodels* provides the answer to both questions. The p-value indicates the strength of evidence for or against H0. A p-value less than or equal to 0.05 suggests a strong significance of the predictor to the response. The coefficient indicates the magnitude of the relationship between the predictor and the response. The larger the coefficient, the larger the change in the response variable for each unit of change in the predictor. It is important to note that the change may be an increase (positive) in the response variable or a decrease (negative).

|  |  |  |
| --- | --- | --- |
| **Predictor** | **Coefficient** | **P-value** |
| **PV\_system\_size\_DC** | 0.6525 | 0.000 |
| **battery\_rated\_capacity\_kW** | 0.2165 | 0.000 |
| **self installed** | -0.1023 | 0.000 |
| **efficiency\_module 1** | 0.0586 | 0.000 |
| **nameplate\_capacity\_module 1** | -0.0096 | 0.006 |
| **multiple\_phase\_system** | -0.0088 | 0.011 |

Table - Predictor Variable Analysis Grouped by Contribution Strength

This table summarizes the results for the predictors studied. All predictors demonstrate statistical significance by having very low p-values. However, the coefficients show that the magnitude of each variable’s contribution varies widely. The team has sorted the predictors by the absolute value of the coefficient. Further, the team grouped the predictors into three groups: those with a high degree of contribution, those with a moderate degree of contribution, and those with minimal contribution. It is important to remember that all variables were standardized prior to modeling. This means all values were transformed to have a mean of 0.0 and a standard deviation of 1.0. Similarly, the coefficients must be understood in this same scale.

Thus, *PV\_system\_size\_DC* is the most influential predictor, contributing a 0.65 increase to the standardized *total\_installed\_price* variable. *Battery\_rated\_capacity\_KW* also contributes considerably to the project price. *Efficiency\_module\_1* makes a moderately positive impact on the project price, while *self\_installed* reduces the price. The remaining two predictors, while statistically significant, do not contribute materially to the *total\_installed\_price*.

The Shapley Additive Explanations (SHAP) method provides a clear and consistent means of visualizing the impact of each feature on the model. (Trevisan, 2022) The SHAP graph below aligns nicely with the team’s analysis of the variable coefficients table. At a glance, viewers can grasp the importance of each predictor in the response variable for each observation.

A screen shot of a graph

Description automatically generated

In summary, the results of this study show that the null hypothesis may be rejected. Several of the selected features statistically impact the total installed cost of residential solar projects in San Diego, California.

One limitation of this study is that no variables that characterize the building upon which the solar project was installed were available. Such variables might include structure square footage, number of floors, and the roof's size, shape, slope, and azimuth. (Office of Energy Efficiency & Renewable Energy, n.d.) These variables would likely influence the cost of a solar project. As a result, the team recommends that stakeholders seek additional data sources with which to augment the inputs to the linear regression model. Doing so could provide additional statistically relevant variables to help build a model for predicting solar project cost.

The team also suggests two additional avenues of study that may benefit stakeholders. First, while this study examined the total project cost, another interesting way of assessing a cost is to determine the dollars per kilowatt generated for the project. This would allow analysts to discover variables that may improve a project's economic efficiency. Second, the study could be widened to include more geographic diversity of observations. This study focused on one city in California. However, there may be broad benefits to having a robust regional or even national solar project cost-predictive model.

# References

Barbose, G., Darghouth, N., Oshaughnessy, E., & Forrester, S. (2024). *Tracking the Sun: Pricing and Design Trends for Distributed Photovoltaic Systems in the United States 2024 Edition*. <https://emp.lbl.gov/sites/default/files/2024-10/Tracking%20the%20Sun%202024_Report.pdf>

Basore, P., Dummit, K., Ellis, C., Peng, P., Benoliel, P., Hendrickson, T., Feldman, D., Ramasamy, V., Woodhouse, M., Zuboy, J., Desai, J., Walker, A., Braid, J., Jost, N., & Cooper, E. (2024). *Solar Photovoltaic System Cost Benchmarks*. Energy.gov. <https://www.energy.gov/eere/solar/solar-photovoltaic-system-cost-benchmarks>

Bobbit, Z. (2019, March 10). *A guide to multicollinearity & vif in regression*. Statology. <https://www.statology.org/multicollinearity-regression/>

Bobbitt, Z. (2020, July 20). *How to Perform a Breusch-Pagan Test in Python*. Statology. <https://www.statology.org/breusch-pagan-test-python/>

Bobbitt, Z. (2022, October 31). *How to Perform Weighted Least Squares Regression in Python*. Statology. <https://www.statology.org/weighted-least-squares-in-python/>

City of San Diego. (2024). *Climate Action Plan 2023 Annual Report* (pp. 5, 8–10). <https://api.watch.kausal.tech/documents/446/2023_CAP_Annual_Report_9-23-24_AM_FINAL.pdf>

Datacamp. (2022). *Python vs R for Data Science:Which Should You Learn?* Datacamp Blog: https://www.datacamp.com/blog/python-vs-r-for-data-science-whats-the-difference

Electricity and Markets Policy Group. (2024, August 21). *Tracking the Sun | Electricity Markets and Policy Group*. Lawrence Berkeley National Laboratory. <https://emp.lbl.gov/tracking-the-sun>

Ellis, C. (2022, May 26). *When to use linear regression*. Crunching the Data. <https://crunchingthedata.com/when-to-use-linear-regression/>

Frost, J. (2019, March 15). *Heteroscedasticity in Regression Analysis*. Statistics by Jim. <https://statisticsbyjim.com/regression/heteroscedasticity-regression/>

Gallo, A. (2016, February 16). *A Refresher on Statistical Significance*. Harvard Business Review. <https://hbr.org/2016/02/a-refresher-on-statistical-significance>

Gulati, A. P. (2022, August 13). *Dealing with outliers using the Z-Score method*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2022/08/dealing-with-outliers-using-the-z-score-method/>

Idriss Jairi. (2024, May 13). *The importance of feature scaling in machine learning: Logistic regression for example*. Medium. <https://medium.com/@jairiidriss/the-importance-of-feature-scaling-in-machine-learning-logistic-regression-for-example-b3068f8d3441>

Jafari, R. (2022). *Hands-On Data Preprocessing in Python* (p. 384). Packt Publishing Ltd. <https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso&db=e000xna&AN=3125175&site=ehost-live&scope=site>

National Renewable Energy Laboratory. (2016). *Solar Energy Evolution and Diffusion Studies: 2014–2016 Customer Motivations and Adoption Barriers Study*. Nrel.gov. <https://www.nrel.gov/solar/market-research-analysis/2014-2016-study.html>

Office of Energy Efficiency & Renewable Energy. (n.d.). *Homeowner’s Guide to Going Solar*. Energy.gov. Retrieved December 14, 2024, from <https://www.energy.gov/eere/solar/homeowners-guide-going-solar>

Plugged In Academy. (2023, June 21). *How Long Does It Take to Install Solar Panels? Timeframe & Factors*. Eco Solar Digest. <https://ecosolardigest.com/how-long-does-it-take-to-install-solar-panels/>

Prashant. (2016, December 28). *How to systematically remove collinear variables (pandas columns) in Python?* Cross Validated. <https://stats.stackexchange.com/questions/155028/how-to-systematically-remove-collinear-variables-pandas-columns-in-python>

Ramasamy, V., Zuboy, J., Oshaughnessy, E., Feldman, D., Desai, J., Woodhouse, M., Basore, P., & Margolis, R. (2022). *U.S. Solar photovoltaic system and energy storage cost benchmarks, with minimum sustainable price analysis: Q1 2022*. <https://www.nrel.gov/docs/fy22osti/83586.pdf>

Srinidhi, S. (2019, November 19). *Backward Elimination for Feature Selection in Machine Learning*. Medium. <https://towardsdatascience.com/backward-elimination-for-feature-selection-in-machine-learning-c6a3a8f8cef4>

Srinidhi, S. (2020, January 20). *Null Hypothesis and the P-Value*. Medium. <https://towardsdatascience.com/null-hypothesis-and-the-p-value-fdc129db6502>

Srinivasan, H. (2020, May 19). *Linear Regression in Scikit-learn vs Statsmodels*. Medium. <https://medium.com/@hsrinivasan2/linear-regression-in-scikit-learn-vs-statsmodels-568b60792991>

Tong, J. (2016, June 20). *How to Turn Solar “Considerers” Into Solar Adopters*. Greentechmedia.com; Greentech Media. <https://www.greentechmedia.com/articles/read/How-to-Turn-Solar-Considerers-Into-Solar-Adopters>

Trevisan, V. (2022, July 5). *Using SHAP Values to Explain How Your Machine Learning Model Works*. Medium. <https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137>

Vijayakumar, A. (2020, November 30). *Stepwise Regression*. GitHub. <https://github.com/AakkashVijayakumar/stepwise-regression>

WGU Information Technology. (n.d.). *R or Python*. https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html

Wu, J., & Coggeshall, S. (2012). *Foundations of predictive analytics*. Crc Press. <https://ebookcentral.proquest.com/lib/westerngovernors-ebooks/detail.action?docID=870690>

Yadav, J. (2023, June 12). *SAS vs R vs Python*. GeeksforGeeks. <https://www.geeksforgeeks.org/sas-vs-r-vs-python/>