**Data Analytics Capstone Topic Approval Form**

**Student Name:** Christopher (Chris) Fischer

**Student ID:** 011933891

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

**Project Topic**:  Identify features of a residential rooftop solar power system that influence the project’s total cost.

**This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research 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?

**Hypothesis**: **Null hypothesis** - The features: system size, single/multi-phase, self-installed, number of modules, module capacity, module efficiency, and battery capacity do not have a statistical impact on the total project cost. **Alternate Hypothesis** - The features: system size, single/multi-phase, self-installed, number of modules, module capacity, module efficiency, and battery capacity do have a statistical impact on the total project cost.

**Context:** 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.

**Data:** The data needed for this study is a set of residential solar projects that include several technical and project attributes and the total project installed cost.

| 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

The data used for this study come from “Tracking the Sun” project run by the Lawrence Berkeley National Laboratory. (Electricity and Markets Policy Group, 2024) The dataset published on the project’s website contains data on over 3 million solar projects. This volume of data exceeds the capability of this researcher to perform studies. Therefore, the dataset will be reduced to include only residential solar projects installed in the greater San Diego, California region during 2023. Since the target variable of this study is price, restricting data to just the most recent year is sensible due to the dual factors marked inflation in the US over the last several years and the ever-downward trend of prices for electronic components. This dataset contains city name and zip code for the location where each project is installed. Since these attributes will not be studied, the data will be further anonymized by setting all city values to “San Diego” and all zip code values to “92101”. The resulting dataset consists of 43,265 observations.

The dataset is owned by the Lawrence Berkeley National Laboratory and is made available for public use without restriction.

**Data Gathering:** The data were gathered by downloading the dataset from this URL: <https://bit.ly/trackingthesun-2024>. (See [this document](https://emp.lbl.gov/sites/default/files/2024-08/LBNL_PublicDataFile_UserGuide_0.pdf) for more detailed information about the dataset.)

**Data Analytics Tools and Techniques**: The primary analysis that will be performed to answer the research question is a multiple linear regression. To get the best results from the regression analysis, additional tools will be used to address the key assumptions made by multiple linear regression models:

* Linear relationship between independent and dependent variables: scatter plots of each independent variable vs. the dependent variable will be visualized to assess the nature of the relationship
* No multicollinearity: variable influence factors will be calculated, and variables with a VIF above a threshold will be removed from consideration
* Homoscedasticity: the resulting model will be examined with a Breusch-Pagan test to detect heteroscedasticity. (Bobbitt, 2020) If detected, the model will be refitted using transformed dependent variable values or a weighted least-squares regression model. (Frost, 2019)

This study will use Python in a Jupyter Notebook with Visual Studio Code as the primary development environment.

The key Python libraries which will be used include:

|  |  |
| --- | --- |
| Module | Purpose |
| Pandas | Provides DataFrame used for data management |
| Matplotlib.pyplot | Tools for visualizing data and results |
| Seaborn | Additional visualization capabilities |
| Statsmodels | Implements many statistical models and tests |

Table - Python Libraries

**Justification of Tools/Techniques:** Multiple linear regression will permit the researcher to determine the direction (increase or decrease) and magnitude to which each independent variable influences the dependent variable. This is accomplished by interpreting the regression coefficients computed by the model. (Sharma, n.d.) Additionally, the statistical significance of each independent variable on the result can be accomplished by evaluating each variable’s p-value. (Gallo, 2016)

As mentioned above, Python was chosen for this study. Other tools were considered, including R and SAS. All three provide outstanding data science capabilities. However, Python was selected because:

* Python is the preferred language for data analysis projects within our organization
* Python has greater flexibility for highly scalable production deployment than R, which is, first and foremost, an interactive analysis tool
* SAS is proprietary, closed source, and very expensive (Yadav, 2023)

Jupyter notebooks are interactive documents that contain code, results, visualizations, and narratives. They provide an excellent means of communicating a study's inputs, processes, and outputs in one convenient vehicle.

**Project Outcomes**: This project will produce a list of residential solar project attributes that statistically impact the cost of deploying a project. Furthermore, it is anticipated that a helpful linear regression model will be produced. This model can be integrated into customer-facing apps or websites for on-demand cost estimates without an initial site visit requirement. The business goals are clear: reduce impediments that keep interested customers from pursuing a solar project. Removing barriers leads to increasing sales. Removing barriers helps San Diego reach its net-zero carbon goals.

**Projected Project End Date**: 12/31/2024

**Sources**:

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

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>

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>

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>

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/>

Sharma, S. (n.d.). Correlation and Regression Analysis: Exploring Relationships in Data. Data Science Salon. Retrieved November 2, 2024, from <https://roundtable.datascience.salon/correlation-and-regression-analysis-exploring-relationships-in-data>

**Course Instructor Signature/Date:**

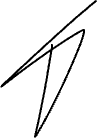
The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Instructor’s Approval Status: Approved

Date: 11/3/2024

Reviewed by:



Comments: Click here to enter text.