FINAL PROJECT REPORT

Solar Power Viability

MIS 6356
Business Analytics With R
Fall – 2020

Prepared By

Igor Rakonjac
Harshit Gupta
Karmani Rajput
Macdonald Ezeruo

NAVEEN JINDAL SCHOOL OF MANAGEMENT
The University of Texas at Dallas

UT DALLAS

Table of Contents

Business Questions and Null Hypothesis

♦ Business Questions/Issues

- The main idea for our project was to explore the viability of solar power and its usability as a means for a long-term future renewable power source.
- Furthermore, we wanted to determine what is the ideal site (state) for a solar panel installation, based on the observations and measurements recorded within the applicable data set, such as Installation Date, PV (photo-voltaic) pricing by system size and across different states, total installed price, appraised value flag, customer segment (pricing for tax-exempt vs. for-profit commercial sites), pricing for residential new construction vs. retrofits, tracking equipment, ground mounted (pricing by mounting configuration), battery system, third party owned, self-installed, sales taxes costs, rebate or grant.
- An example that can be utilized as an opportunistic idea to support solar power promotion, is the amount of sunlight available in Texas. It can be represented by the following, which in turn classifies Texas as one of the top candidates to be a "Sunlight" state (it is also the main driving factor for the undertaking of this project):
  - % of Sun: 61, which represents the percentage of time between sunrise and sunset when sunshine reaches the ground.
  - Total Hours: 2850, which is the average number of sunny hours that a given place in Texas normally has in a year.
  - Clear Days: 135, which is the average number of days annually when clouds cover at most 30% of the sky during daylight hours.
  - Average temperature of >82°F in the summer months (5+ months).

♦ Null Hypothesis

- **Installation cost for solar plants varies for different customer segments.**

Data Set Description

♦ Availability on Public Domain/Source of Data Set:

- The data is available on public domain and maintained by US Department of Energy and National Renewable Energy Laboratory (NREL). The data is related to installation cost for various sites in collaboration with Lawrence Berkeley National Laboratory.
- Other than the data set we found, there are other articles and resources available online regarding the development and growth of renewable energy industry across the United States, as well as globally. We have utilized this information to support

our research, determine the validity of our Null Hypothesis and arrive at a conclusion. See References section for more details.

♦ Basics of Data:

- Data set provides installed prices and other trends among grid-connected, distributed solar photo-voltaic systems across different states in the United States.
- The available data covers the period between 1998 and 2018, with preliminary trends for the first half of 2019.
- Data set includes only grid-connected systems, mainly defined as rooftop, or ground mounted systems, up to 5MW (AC).
- Data set includes data for more than 1.5 million PV systems installed by the end of 2018, with over 60 data fields describing key attributes of each system.
- The observations are stored in two separate ".csv" files, which were combined using RStudio for the purpose of our analysis.

♦ Major Data Attributes:

The main data variables are explained in Table-1. After reviewing and analyzing the data available, we are utilizing a total of 14 variables we believe would best assist us in determining the validity of our Null Hypothesis.

Table-1: Description of Data Variables

Data Field Name	Units	Description of Data Field	Data Type
Installation Date	date	For some data providers, the installation date may be based on the best available proxy, such as the date that an incentive claim was submitted or when the inspection was performed.	Char
System Size	kW (DC)	The total rated direct-current (DC) output of the module arrays at standard test conditions. These data are generally reported directly by the data provider, but in some cases must be estimated, for example, based on the module model and quantity or based on reported alternating-current (AC) capacity.	Num
Total Installed Price	$ (nominal)	The total installed price for the system, prior to receipt of any incentives, as reported by the installer, host customer, or other incentive applicant. For third-party owned systems, the data may represent one of two things. If the third-party owner procured the system from an independent installation contractor, then the reported installed price likely refers to the	Num

		intermediate sale price between the installation contractor and the third-party owner. If the third-party owner instead installed the system itself, then the reported installed price likely represents an appraised value. The installed price data may be subject to any number of other reporting inconsistencies, which may or may not be readily identifiable. In addition, the data may suffer simply from self-reporting errors, and the level of verification vary across data providers.	
Appraised Value Flag	N/A	A flag used to indicate whether the reported installed price is likely to represent an appraised value. Caution should be used in relying on appraised values for analysis or benchmarking purposes, as such data do not represent a transaction price.	Log
Customer Segment	N/A	Data on customer segment is mapped to one of six general types: RES, COM, SCHOOL, GOV, NON-PROFIT, and NONRES, the last one being used only if more-specific information on non-residential customer type is unavailable.	Char
New Construction	N/A	Indicates if the system was installed at the time of building construction. Data generally available for only those states or utilities that have separate programs or incentive rates for new construction vs. retrofits.	Int
Tracking	N/A	Indicates if the system includes tracking equipment.	Int
Ground Mounted	N/A	Indicates if the system is ground-mounted (which may include pole-mounted systems). PV systems consisting of a combination of rooftop and ground-mounted arrays are coded as ground-mounted.	Int
Battery System	N/A	Indicates if the system includes batteries.	Int
Third-Party Owned	N/A	Indicates if the system is third-party owned; that is, owned by an entity other than the site host and either leased or sold under a power purchase agreement to the site host.	Int

Self-Installed	N/A	Indicates if the system was installed by the site-host.	Int
Sales Tax Cost	$ (nominal)	The calculated cost of sales taxes. This is estimated based on average sales tax rates for the given state and year, accounting for any sales tax exemptions that may exist for PV systems. Sales taxes, if applicable, are assumed to be levied only on hardware costs, which are assumed to represent 55% of the total installed price.	Num
Rebate or Grant	$ (nominal)	The pre-tax value of any up-front rebate or grant provided by the entity supplying the data.	Num
State	N/A	Host customer state.	Char

Explanatory Data Analysis

- **Overall Data Summary:**
  - Two ".csv" files joined before conducting the analysis.
  - As per the data source recommendation, any missing values within the data set(s) have been replaced with '-9999' value(s).
  - From Figure-1 it is clear that data has more than 1.5 million observations and 60 variables. All these variables are classified as character, numeric or logical type of data.

```
Name                        data.df
Number of rows              1543831
Number of columns           60
----------------------
Column type frequency:
  character                 26
  logical                   1
  numeric                   33
```

Figure-1 Data Summary for Original Data

  - "str()" function was used to see type of class for each variable and type of observations.
  - From reviewing the output, it is clear that we require data selection and data cleaning for implementation of effective predictive model.
  - Output is too large to include in report hence viewed in knit.pdf file.

- **Selection of Variables:**
  - Out of 60 variables present within the data sets some of the variables have 50% or more number of observations are missing values. Hence those variables have not been considered for further analysis. Some of such variables are Azimuth, Tilt, BIPV Module, Module Efficiency, etc.
  - Some of variables have P25= 0 or -9999 and P100 =1. These variables are more suited as Factor type. However, they were analyzed as numeric data points only.
  - As per Null Hypothesis, Total Installed Price for the PV systems would be predicted based on selected variables. Thus, the selected variables for the analysis are Installation Date, System Size, Total Installed Price, Appraised Value Flag, Customer Segment, New Construction, Ground Mounted, Tracking, Battery System, Third-Party Owned, Self-Installed, Sales Tax Cost, Rebate/Grant, State.
  - Summary for selected data set is shown in Figure-2 and details for each type of variables are also shown in Figure-3 and Figure-4 for character type and numeric type of data respectively.

Name	data_na.df
Number of rows	1187660
Number of columns	18
Column type frequency:	
character	7
logical	1
numeric	10

Figure-2 Data Summary for Selected Data

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Installation.Date	0	1.00	8	10	0	6104	0
Customer.Segment	24250	0.98	3	10	0	6	0
State	0	1.00	2	2	0	24	0
Installer.Name	109090	0.91	1	66	0	11460	0
Module.Technology..1	175796	0.85	3	14	0	11	0
Module.Technology..2	1135985	0.04	3	14	0	11	0
Module.Technology..3	1172889	0.01	4	11	0	9	0

Figure-3 Data Summary for Character Type Variables

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
System.Size	0	1.00	11.21	76.61	0.00	4.22	6.00	8.41	10000	
Total.Installed.Price	0	1.00	50564.41	4448807.25	0.01	18270.00	26433.00	37440.00	4806584994	
Sales.Tax.Cost	195	1.00	889.58	7978.94	0.00	0.00	360.15	642.43	1735160	
Rebate.or.Grant	89336	0.92	3686.92	32337.18	0.00	0.00	0.00	2402.00	5500000	
New.Construction	459005	0.61	0.07	0.25	0.00	0.00	0.00	0.00	1	
Tracking	355624	0.70	0.01	0.08	0.00	0.00	0.00	0.00	1	
Third.Party.Owned	102732	0.91	0.43	0.49	0.00	0.00	0.00	1.00	1	
Self.Installed	204171	0.83	0.02	0.14	0.00	0.00	0.00	0.00	1	
Ground.Mounted	506633	0.57	0.03	0.17	0.00	0.00	0.00	0.00	1	
Battery.System	400119	0.66	0.01	0.12	0.00	0.00	0.00	0.00	1	

Figure-4 Data Summary for Numeric/Int Type Variables

Data Processing

- **Missing Values and Outliers:**
  - The Data containing missing values has been replaced with "-9999".
  - While conducting our data cleaning phase, we have noticed that the presence of dummy values in place of missing values, created a challenge of getting actual statistics of variables. Hence, to avoid false analysis we have replaced these values with "NA" values.
  - We have noticed that the actual installation price data contains 356171 null values and 63336 outliers, whose value is much higher than the upper control limit of all data. As we are analyzing installation price, we have removed the observation where the installation price has any null values. However, we are keeping the outliers for our analysis, as we feel that removing those outliers might result in false prediction(s). Therefore, the box plot has only one visible line (see Figure-A1). Similarly, if we remove the outliers, we can see a more visible box plot. (see Figure-A2) (FigureA1 and A2 are in Appendix-A)

- **Installation Price – Customer Segment:**
  - After selecting 14 predictors, Average Installation Price is compared over the various customer categories. See Figure-5 and Table-2.

Table-2: Average Installation Price – Customer Segment

	Customer_Segment	Average_Price
1	COM	524197.09
2	GOV	890683.31
3	NON-PROFIT	230882.56
4	NON-RES	297257.60
5	RES	34382.72
6	SCHOOL	738943.93

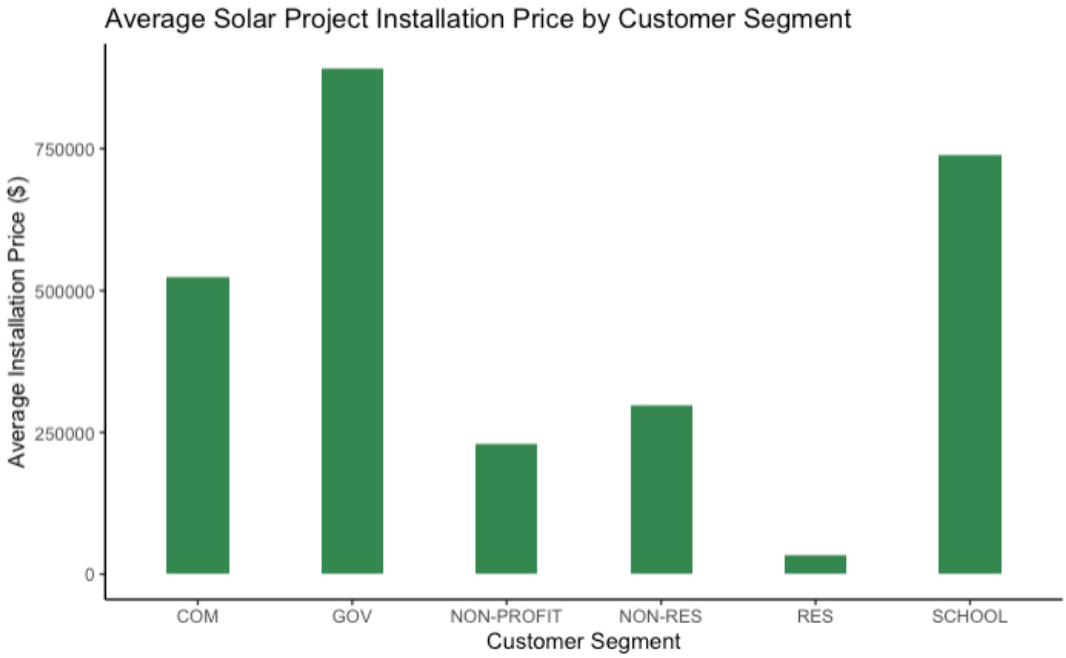

Figure-5 Average Installation Price – Customer Segment

♦ Data Cleaning:

- We have removed all the observations which have "NA" values from all the variables, so that our model is our more accurate.
- In the given data set(s), installation date had two different formats of representation, one with "MM/DD/YYYY" and another as "MM-DD-YYYY".
- To unionize the data, first we have replaced all dates in mm-dd-yyyy format and then split the data into separate columns as "MM", "DD" and "YYYY".
- For Appraised Valued Flag, which is a logistic variable, we have converted TRUE into "1" and FALSE into "0".
- For Customer Segment predictor, we have assigned the different categories into 1-6 numbers, as shown in the Table-3 below:

Table-3: Average Installation Price – Customer Segment

Customer Segment	Numerical Representation
Non-Residential	1
Residential	2
Government	3
School	4
Commercial	5
Non-Profit	6

- Variables type are changed as described above. See Table-4 for more details regarding the clean data through str() function. Output for the same is shown in Appendix-B

Table-2: Description of Cleaned Data Variables

Data Field Name	Data Type	Cleaned Data Type
Installation Date (split into 3 subcategories below)	Char	Int
Installation.Data.1 = Month		
Installation.Data.2 = Day		
Installation.Data.3 = Year		
System Size	Num	Num
Total Installed Price	Num	Num
Appraised Value Flag	Log	Num
Customer Segment	Char	Num
New Construction	Int	Int
Tracking	Int	Int
Ground Mounted	Int	Int
Battery System	Int	Int
Third-Party Owned	Int	Int

Self-Installed	Int	Int
Sales Tax Cost	Num	Num
Rebate or Grant	Num	Num
State	Char	Char

Empirical Data Analysis

♦ Multiple Linear Regression

- As we have both categorical and numerical data types, and our model choice depends on our target variable (pricing), we have determined that due to it being continuous, Multiple Linear Regression approach was utilized.

- As we have converted the categorical variables to numeric variables, after running the Multiple Linear Regression all other variables are significant except for "Tracking"

- While we were testing the accuracy of the model (in validation data), we were able to reduce the RMSE (Root Mean Square Error), which means we have a much lesser error in the accuracy of our model.

- After running the regression (see Figure-6 below), we obtained the following results:

  Residual standard error: 14720 on 427199 degrees of freedom; a square root of the residual sum of squares divided by the residual degrees of freedom, which shows us a good fit via our linear regression model.

  Multiple R-squared: 0.7129; a statistical measured of how close the data are to our fitted regression line. As we are relatively close to 100% and in conjunction with the residual plot, our model successfully explains the variability of the response data around its mean.

  Adjusted R-squared: 0.7129; a comparator of the explanatory power of regression models that contain different number of predictors. Similarly, as with the multiple R-squared, as our model has virtually no non-significant variables, there is no gap between two.

  F-statistic: 8.84e+04 on 12 and 427199 DF; a value we get when we ran a regression analysis which shows that it provides a good fit to the data.

  p-value: < 2.2e-16; as the p-value is less than the significance level, our sample data provides sufficient evidence to conclude that our regression model first the data better than say a model with no independent variables.

```
    Min      1Q  Median      3Q     Max
-84.786  -0.156  -0.020   0.148 104.857

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.145e-14  8.148e-04   0.000   1.0000
Installation_Month   -1.421e-02  8.170e-04 -17.397   <2e-16 ***
Installation_Year     7.049e-02  9.760e-04  72.228   <2e-16 ***
System.Size           2.943e-01  1.322e-03 222.663   <2e-16 ***
Sales.Tax.Cost        5.076e-01  1.515e-03 334.998   <2e-16 ***
Rebate.or.Grant       1.576e-01  1.010e-03 156.016   <2e-16 ***
Customer.Segment     -1.815e-02  8.173e-04 -22.202   <2e-16 ***
New.Construction     -4.641e-02  8.396e-04 -55.270   <2e-16 ***
Tracking             -1.550e-03  8.164e-04  -1.898   0.0577 .
Appraised.Value.Flag  1.013e-01  8.924e-04 113.472   <2e-16 ***
Third.Party.Owned    -5.817e-02  8.942e-04 -65.056   <2e-16 ***
Ground.Mounted        1.827e-02  8.276e-04  22.079   <2e-16 ***
Battery.System       -4.339e-02  8.228e-04 -52.737   <2e-16 ***
Self.Installed       -3.772e-02  8.227e-04 -45.844   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5326 on 427198 degrees of freedom
Multiple R-squared:  0.7164, Adjusted R-squared:  0.7164
F-statistic: 8.3e+04 on 13 and 427198 DF,  p-value: < 2.2e-16
```

Figure-6: Multiple Linear Regression Outputs

♦ Ridge and Lasso Regression

- After running the ridge regression (see Figure-7 below), we obtained the following results:

```
Resampling results across tuning parameters:

  lambda    RMSE       Rsquared   MAE
  0.000100  0.5661909  0.6840889  0.2321530
  0.250075  0.5623917  0.6899127  0.2395345
  0.500050  0.5727698  0.6914021  0.2515130
  0.750025  0.5893555  0.6911202  0.2627119
  1.000000  0.6077200  0.6904307  0.2727575

Tuning parameter 'alpha' was held constant at a value of 0
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 0 and lambda = 0.250075.
```

Figure-7: Ridge Regression Outputs

- After running the lasso regression (see Figure-8 below), we obtained the following results:

```
Resampling results across tuning parameters:

  lambda      RMSE       Rsquared   MAE
  0.00010000  0.5608799  0.6940474  0.2302283
  0.03342222  0.5678021  0.6864005  0.2377812
  0.06674444  0.5786567  0.6771553  0.2495095
  0.10006667  0.5885492  0.6710239  0.2563024
  0.13338889  0.5977653  0.6677306  0.2596131
  0.16671111  0.6084028  0.6646077  0.2635947
  0.20003333  0.6205438  0.6615302  0.2682478
  0.23335556  0.6340062  0.6585473  0.2736155
  0.26667778  0.6486526  0.6555987  0.2795941
  0.30000000  0.6646892  0.6523531  0.2861135

Tuning parameter 'alpha' was held constant at a value of 1
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were alpha = 1 and lambda = 1e-04.
```

Figure-8: Lasso Regression Outputs

Conclusion

- Based on our analysis, as the Customer Segment and Sales Tax Cost are significant to Total Installation Price of Solar Panel(s), we can say with certainty that we have enough evidence to support out Null Hypothesis.
- From figure-6 to 8 we can say that multiple linear regression has highest R squared value hence we predicted our validation data by using that model which output is shown in Fig.-9 and Fig.-10.

```
##                     ME      RMSE       MAE       MPE     MAPE
## Test set -0.002341334 0.4440154 0.2300935 6.974286 249.1205
```

Figure-9: Accuracy Output for Multiple Linear Regression

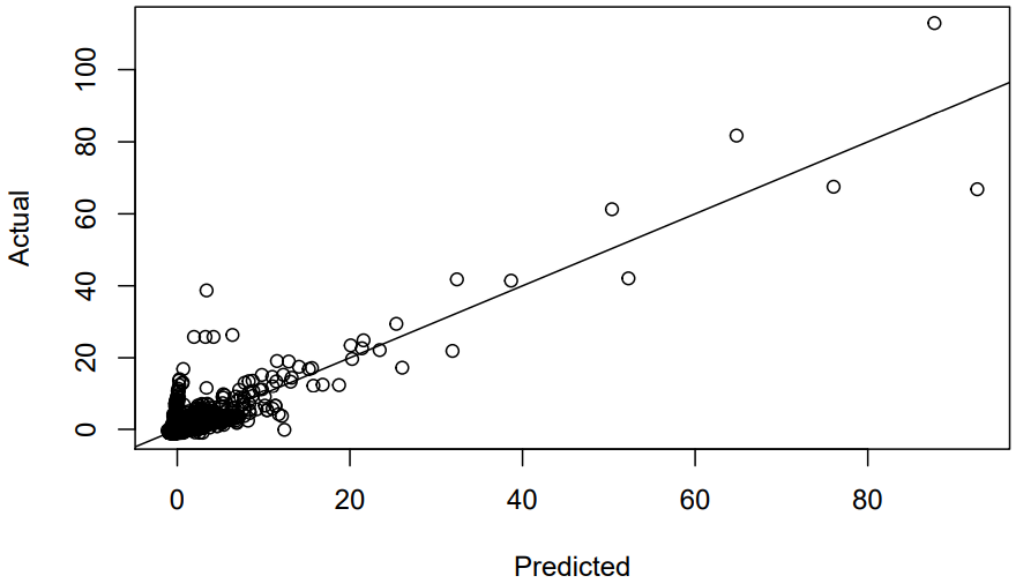

Figure-10: Actual-Predicted Output for Multiple Linear Regression

APPENDIX – A

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000e+00	1.827e+04	2.643e+04	5.056e+04	3.744e+04	4.807e+09

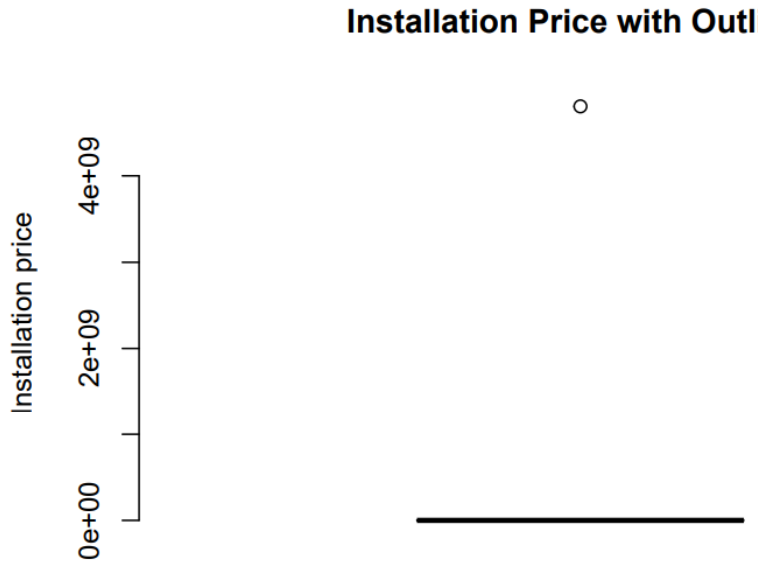

Figure-A1: 5-Number Summary and Box Plot for Total Installation Price (with Outliers)

LCL	Q1	Median	Mean	Q3	UCL
-10485.00	18270.00	26433.00	50564.41	37440.00	66195.00

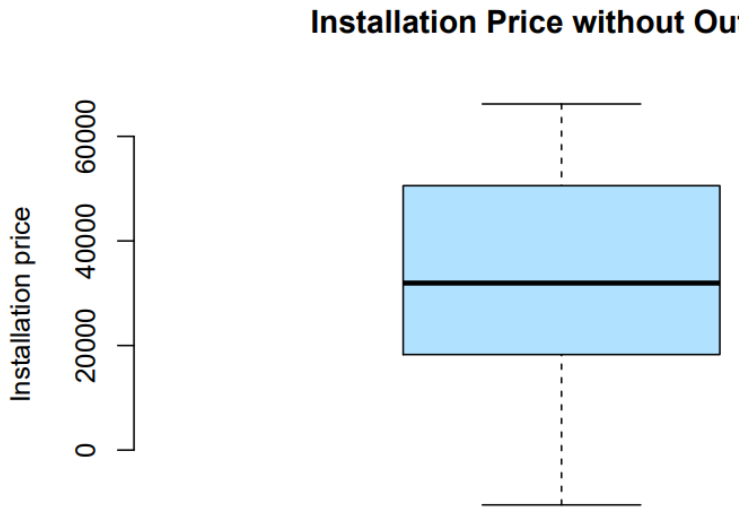

Figure-A2: 5-Number Summary and Box Plot for Total Installation Price (without Outliers)

APPENDIX-B

```
str(Clean_data.df)
```

```
## Classes 'data.table' and 'data.frame':   534012 obs. of  14 variables:
##  $ Installation_Month   : num  6 2 9 8 1 12 9 10 10 3 ...
##  $ Installation_Year    : num  2010 2010 2010 2011 2010 ...
##  $ System.Size          : num  4.42 2.99 2.99 2.99 2.99 4.42 3.04 5.33 2.99 2.99 ...
##  $ Total.Installed.Price: num  26862 20884 18877 18305 20884 ...
##  $ Sales.Tax.Cost       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Rebate.or.Grant      : num  11934 8073 6428 4036 8073 ...
##  $ Customer.Segment     : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ New.Construction     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Tracking             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Appraised.Value.Flag : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Third.Party.Owned    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Ground.Mounted       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Battery.System       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Self.Installed       : int  0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Figure-B1 Summary for Clean Data

References

1. “Tracking the Sun.” Tracking the Sun | Electricity Markets and Policy Group, emp.lbl.gov/tracking-the-sun.

2. “Which States Are Best for Solar Power?” Which States Are Best for Solar Power? | Vivint Solar Learning Center, www.vivintsolar.com/learning-center/top-states-for-solar.

3. “How Much Do Solar Panels Cost in Texas? See Prices, Rebates & Tax Credits.” Solar Reviews, 7 Nov. 2020, www.solarreviews.com/blog/how-much-do-solar-panels-cost-in-texas.

4. “Solar Power Incentives In Texas.” DIY Solar For Your Home, 29 June 2015, www.mydiysolarhouse.com/solar-power-incentives-in-texas/.

5. “Days of Sunshine Per Year in Texas.” Annual Days of Sunshine in Texas - Current Results, www.currentresults.com/Weather/Texas/annual-days-of-sunshine.php.