Abstract:

Tankus Industries (TI) was contacted to analyze Oregon housing data from the 2015 American Community Survey. The main task of this analysis is to determine how expensive electricity bills in Oregon are, and what factors contribute to their cost. TI was asked to pay close attention to the relationship between electricity cost of apartments versus houses, and to adjust for the number of bedrooms and occupants in a house. We are then asked to create a model to predict electricity costs for a typical Oregon household. TI suggests breaking these into smaller, more manageable questions to aid in analysis as follows:

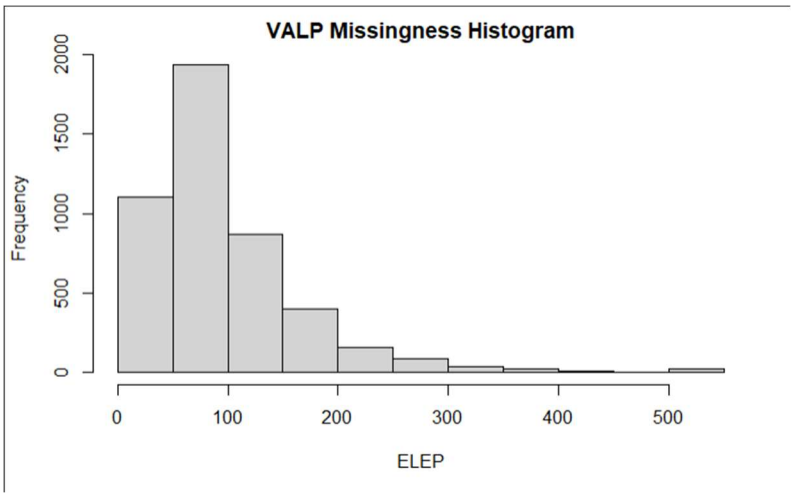
1. Does the data require any cleaning?
2. How much do people pay for electricity in apartments?
3. How much do people pay for electricity in houses?
4. Is there a statistical difference between 2 and 3?
5. What model best predicts electricity cost?

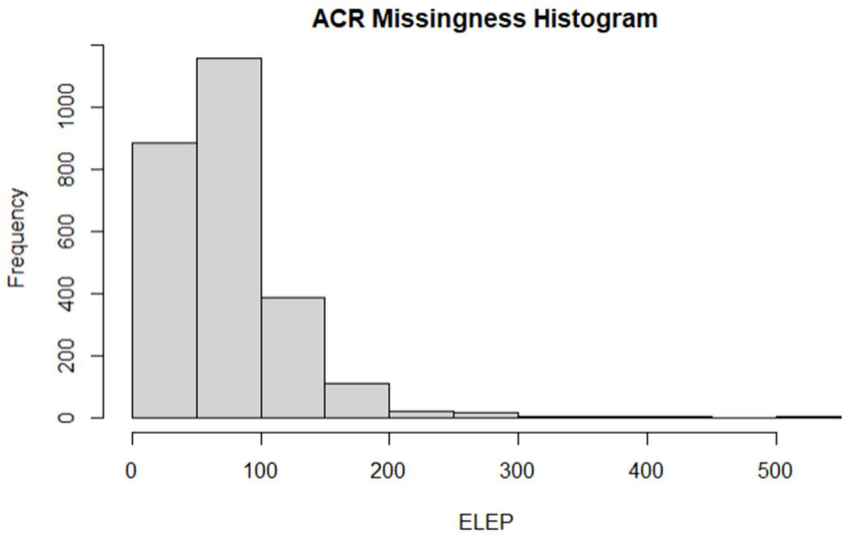
Contents

Does the Data Require Any Cleaning?

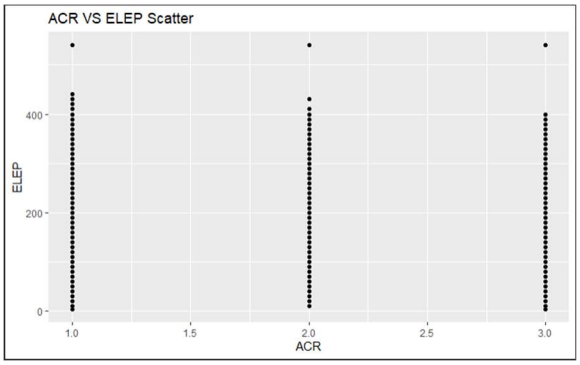
As with most data, this set did require cleaning. Mostly regarding the missing values in the data. ACR and VALP both have many null values, 2586 and 4632 respectively. Looking at the missingness distribution histograms, it seems both VALP and ACR are missing-not-at-random, and clearly have more missing values at the low end of the ELEP histograms.

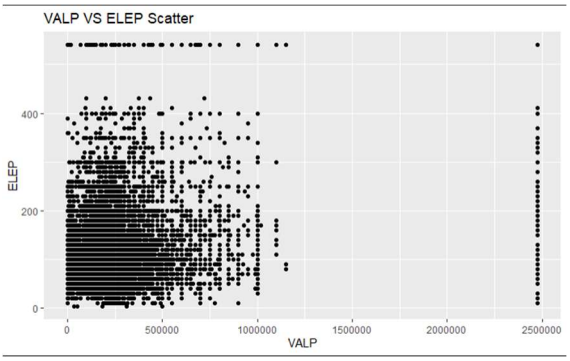
TI believes that these two variables should be removed from the analysis, as property value (VALP) and Lot Size (ACR)





The scatter plots of ACR/VALP VS ELEP also show minimal correlation, meaning it is unlikely that removing ACR and VALP will have a negative impact on the study.





Looking at the missingness identified the large range of electricity costs in the sample. This led TI to a histogram of the monthly electricity cost (ELEP).

Choosing Practically Relevant Fields

Once the missingness of the data was analyzed, TI took the steps to rank relevant fields, and potentially remove any practically irrelevant fields from our data. Once identified, these fields will have a correlation check with ELEP before complete removal.

Some categorical fields (BLD, HFL, TEN, YBL, R18, R60) needed converting to numeric for correlation consideration. The full scatter matrix is available IN APPENDIX X, and correlation metrics are listed below. None of the practically relevant factors have noteworthy correlation, so TI recommends removing them from the model. RMSP and BDSP are also highly correlated (0.72), so TI recommends removing BDSP to prevent redundant information. FULP, GASP, and YBL may still be relevant after further inspection, so TI recommends leaving them in along with BLD, HFL, RMSP, NP, and ELEP.

Field	Description	Practically Relevant to ELEP?	Correlation to ELEP
SERIALNO	Serial Number	No	N/A
TYPE	Type of Unit	No (always 1)	N/A
R18	Presence of persons under 18	No	0.16
R60	Presence of persons over 59	No	0.013
TEN	Tenure	No	0.077
FULP	Yearly Fuel Cost (other than gas & electricity)	Maybe	0.06
GASP	Gas (monthly cost)	Maybe	0.029
YBL	When structure first built	Maybe	0.01
NP	Number of Persons in the house	Yes (not in current form)	0.28
BLD	Units in Structure	Yes (not in current form)	0.18
HFL	House Heating Fuel	Yes (not in current form)	0.14
RMSP	Number of Rooms	Yes	0.23
BDSP	Number of Bedrooms	Yes (Redundant)	0.26
ELEP	Electricity (monthly Cost)	Yes	1

Cleaning Data and Transitioning Relevant Data to a Computer Digestible Format

Now that the practically relevant fields have been identified, TI must convert them to more useable data, based on the research question. For example, there are many housing types listed under BLD, but

the research question only asks for differences between apartments and houses. TI recommends analyzing these data as “Apartment” and “House” values, as shown in the table below. HFL also has many types of fuel and TI similarly recommends using “Electricity” and “Not Electricity” values for this field. TI also recommends grouping houses newer than 2005 into one “2005 to 2015” group to reduce model complexity.

Field	Previous Value	Analyzed Value
BLD	Mobile home or trailer	Removed
BLD	Boat, RV, van, etc.	Removed
BLD	One-family house detached	House
BLD	One-family house attached	House
BLD	2 Apartments	Apartment
BLD	All other Apartment Fields	Apartment
HFL	Electricity	Electricity
HFL	All other non-electricity fields	Not Electricity
YBL	Year-by-Year for 2005+	2005 to 2015

Do People Living in Apartments Pay Less on Electricity than those Living in Houses?

This is the main question TI was asked, and we are now able to answer it. TI has analyzed the data by fitting two separate models, one for each House and Apartment dwelling types. TI recommends iteratively fitting and analyzing these two models with the potentially relevant fields until the optimal balance between model complexity and model fit is achieved. The best explanatory variables to include in both models are HFL, NP, and RMSP which yields a R^2 around 0.76 in each model. In both models NP and RMSP are both large enough to be treated as fixed effect, continuous variables while HFL will be analyzed as a categorical “Electricity” and “Not Electricity” values. Once the fitted values of these two models are compared and TI found **electricity is more expensive in a house than in an apartment** with median values at $121.60 and $81.17, respectively. This is roughly a **$40 per month difference**.

Creating an Electricity Model to Predict Electricity Costs in Oregon

Creating a model to *predict* electricity costs is more difficult than just fitting the current data. To fit the data TI plans to fit models using the forwards and exhaustive validation set approaches with a 10-fold k-means cross-validation to ensure valid model comparison. Mean Squared Error, Adjusted R-Squared, BIC, and CP values will be analyzed for each model and compared. TI will then recommend either the exhaustive or forwards validation methods based off these four-comparison metrics.