UNIVERSITÁ DEGLI STUDI DI MILANO

STATISTICAL LEARNING
M.SC. DATA SCIENCE & ECONOMICS

Mini Excavators Market Size in Turkey: a Statistical Learning approach

David Heilbron
Code: 988346

September 18, 2022

Contents

List of Tables

List of Figures

Introduction

Excavators are specialized equipment used in the construction sector across different types of construction, and could be considered one of the most elemental or must-have equipment for almost any kind of project: from digging a pool in a backyard to dreading a canal or move massive amounts of debris in a demolition site. This report is then focused on estimating a certain type of excavator called mini excavator. These are the lightest type of the excavators, versatile and practical, they are designed for different niche markets than the traditional excavators (Figure 1).

With this in mind, as in any other business knowing the size of a market is important. On the one hand, it's useful to design strategies, size opportunities and track competition performance. On the other hand, properly estimating it requires a handful of resources and knowledge. Specially with B2B (Business to Business) markets, finding reliable information is quite a hard task, and when found, it is usually only available for developed or high-income countries, as developing countries do not rely on solid statistical institutions and measurement systems. This is the case for Turkey, for which even specialised market research companies do not have proper estimates of the number of excavators sold in a year.

The motivation behind this project is to give an answer to the business question: *how many mini excavators were sold in Turkey in 2021?* Even though the goal is to estimate an exact number, building a path and methodology to apply statistical and machine learning techniques to real world business problems could be considered as equally important.

Figure 1: Mini Excavator in use

Data and Strategy

Unsupervised Learning

The strategy to answer the business question is divided in two phases. Phase one aims to find similar characteristics between the Turkish construction market and other countries. For this purpose, **Unsupervised Learning** techniques such as *Principal Component Analysis*, *Hierarchical Clustering* and *K-means Clustering* are used. Data was scaled to have a zero mean and standard deviation of 1 before applying the algorithms.

For this first first stage, techniques are applied to a database called "Equipment". The data is composed by the market value of all the equipment sold to new constructions in 2021 for 16 countries, Turkey among them. The values are distributed in 24 end sector categories, ranging from Rail Infrastructure Equipment to Single-Family housing Equipment. Values are expressed in real USD millions. Source is Global Data, a market research company.

Supervised Learning

For the second phase of the project, focused on **Supervised Learning** techniques, the goal is to estimate the number of Mini Excavators sold in Turkey in 2021. Available information is the dependent variable, number of Mini Excavators sold in 15 different countries (Turkey not available) in 2021 and 323 different economical and demographic features for each country for the same year. Features are numerical values expressed either on USD currency values or index value, displayed as a percentage. Data with an absolute skewness value higher than 1 was log transformed, then, all features were scaled.

The countries used are the same as in the Unsupervised Learning phase, but without Turkey. Source of the dependent variable is Off Highway Research (OHR), a specialized market research company focused on the construction sector, and for the independent variables the source is again Global Data.

The reader may be asking herself why did I selected only 15 countries to make the estimation. The reason is that that's all the available information. It is a clear limitation on the number of techniques that can be applied, but as mentioned before, the goal of the project is to solve a problem in a real business scenario.

The techniques used in the second phase are linear regressions with regularization: *Lasso*, *Ridge* and *Elastic Net*. Leave One Out Cross Validation (LOOCV) is used on the three techniques to estimate the lambda with the lowest Sum of Squared Errors (SSE). For the Elastic Net regularization, the best lambda was chosen first and then the best alpha among ten different values ranging from 0.1 to 1, using again LOOCV.

The best model is chosen based on the lowest LOOCV SSE, then a Bootstrap Regression is implemented on the chosen model to accurately determine the standard errors and significance of the features. The best model is then used to estimate the number of Mini Excavators sold in Turkey in 2021.

Results

Unsupervised Learning

Principal Component Analysis

Principal Component Analysis was run over the "Equipment" dataset with the objective of reducing the dataset and conduct feature analysis. The Scree Diagram on Figure 2 shows the explanatory power of the first ten Principal Components (PC). It is clearly visible that the first two PC encompass most of the variation, and the third component represent as much variation as a single feature.

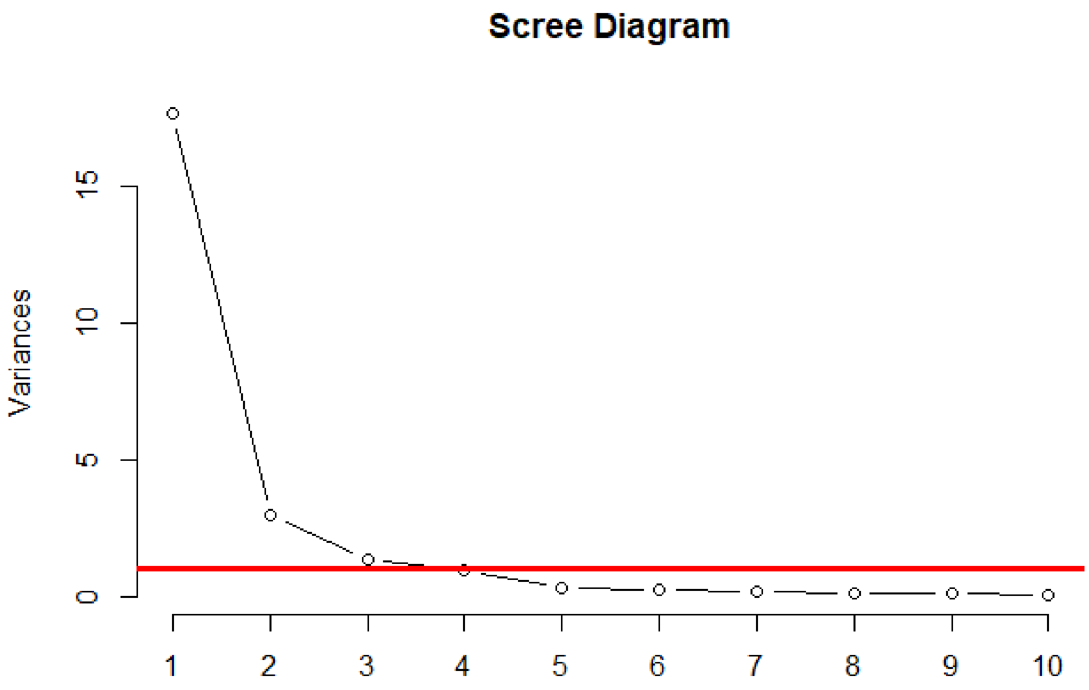

Figure 2: Scree Diagram, Principal Components of Equipment dataset

With this in mind, a biplot of PC1 and PC2 shows the direction in which the variables move relative to the components and how the countries are grouped in relation with them (Figure 3) A simplified graph showing only the countries position on the plane can be observed on Figure 4, where a cluster is clearly identified as well as some countries that show completely different characteristics (Germany, France, United Kingdom).

If the distance between Turkey and the rest of the countries is measured for each component, for PC1, the Netherlands is the closest one and for PC2 is Italy, suggesting similar dynamics between their equipment markets.

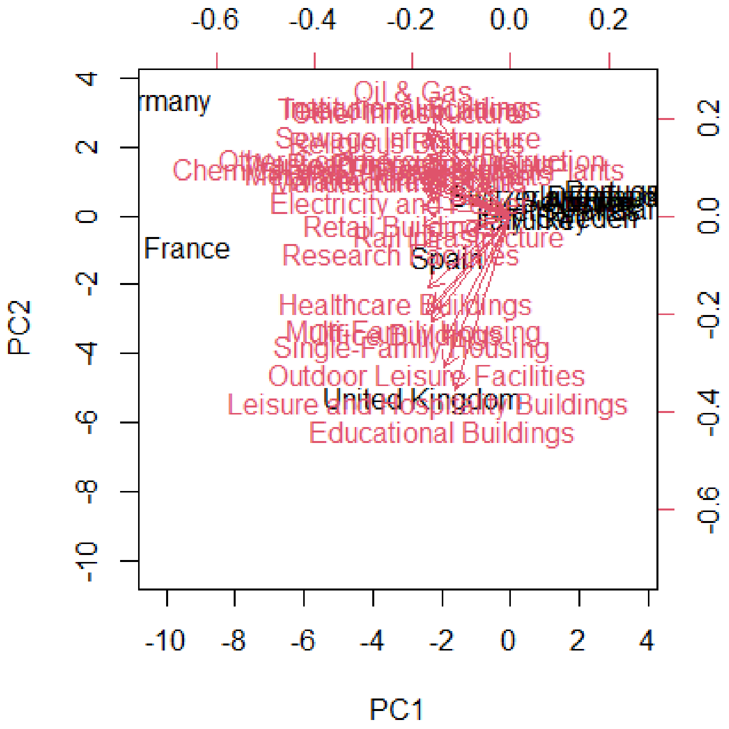

Figure 3: Biplot, PC1 and PC2 of Equipment dataset

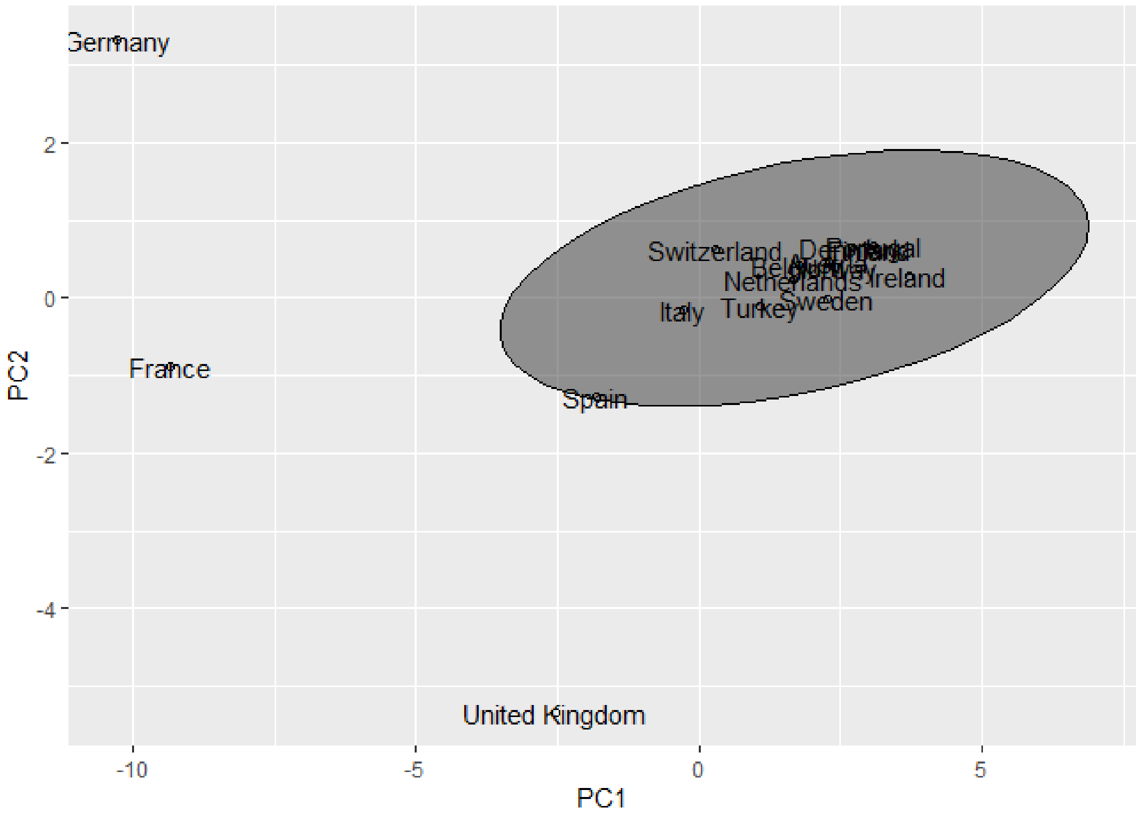

Figure 4: Scatterplot, PC1 and PC2 of Equipment dataset

Hierarchical Clustering

To test another technique and validate results, Hierarchical Clustering is applied to the data. Euclidean distance is applied as a distance measurement and three different linkage methods are tested: Complete Linkage, Average Linkage and Ward Distance. General results are similar to the ones of PC analysis but with the small difference that Turkey resembles more to Switzerland than to any other country in all three methods tested (Figure 5).

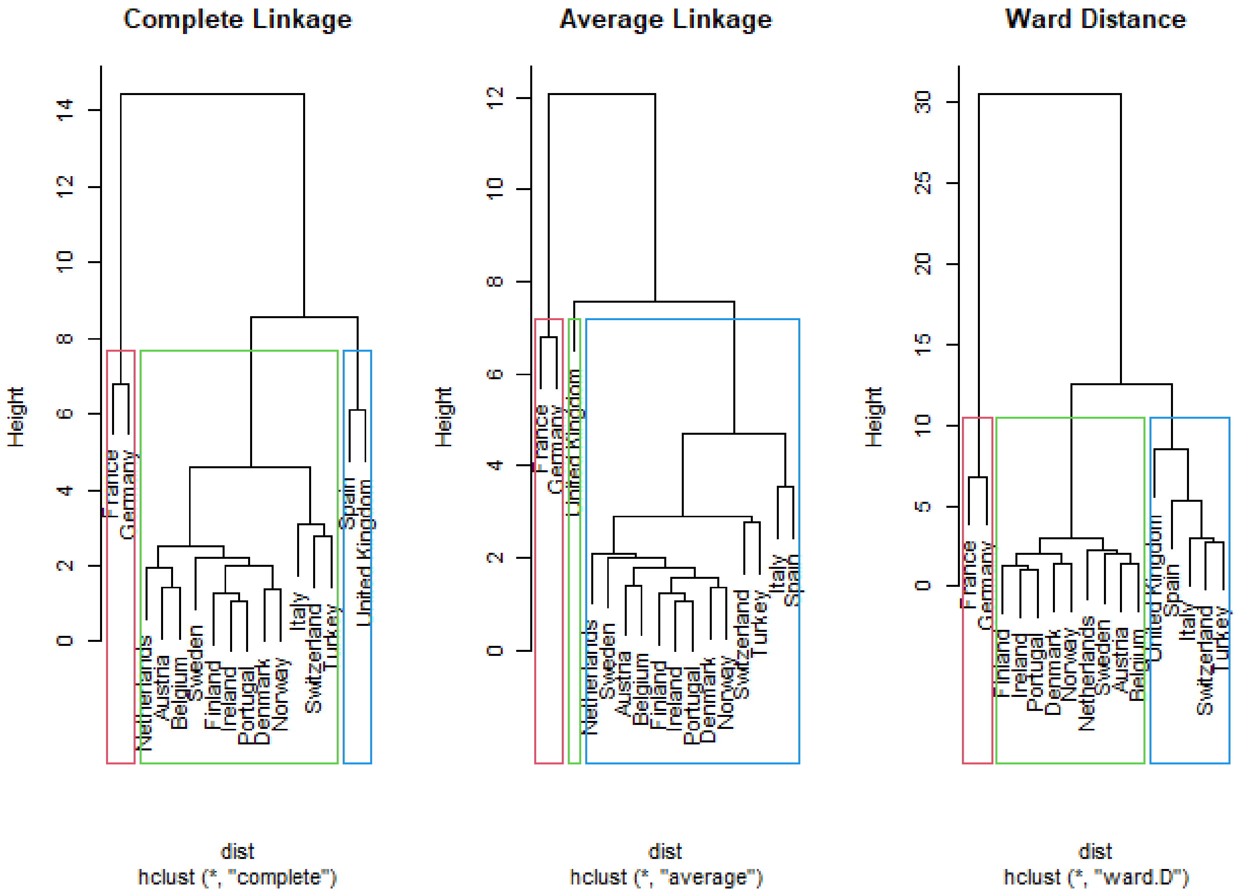

Figure 5: Hierarchical Clustering approaches

To understand what does it mean to belong to a certain cluster, a scatter plot diagram of some variables are shown in Figures 6 and 7, for this example only the complete linkage clustering is shown. Here, it is clear how the different markets are split according to the size of their end sectors and correlations among each other.

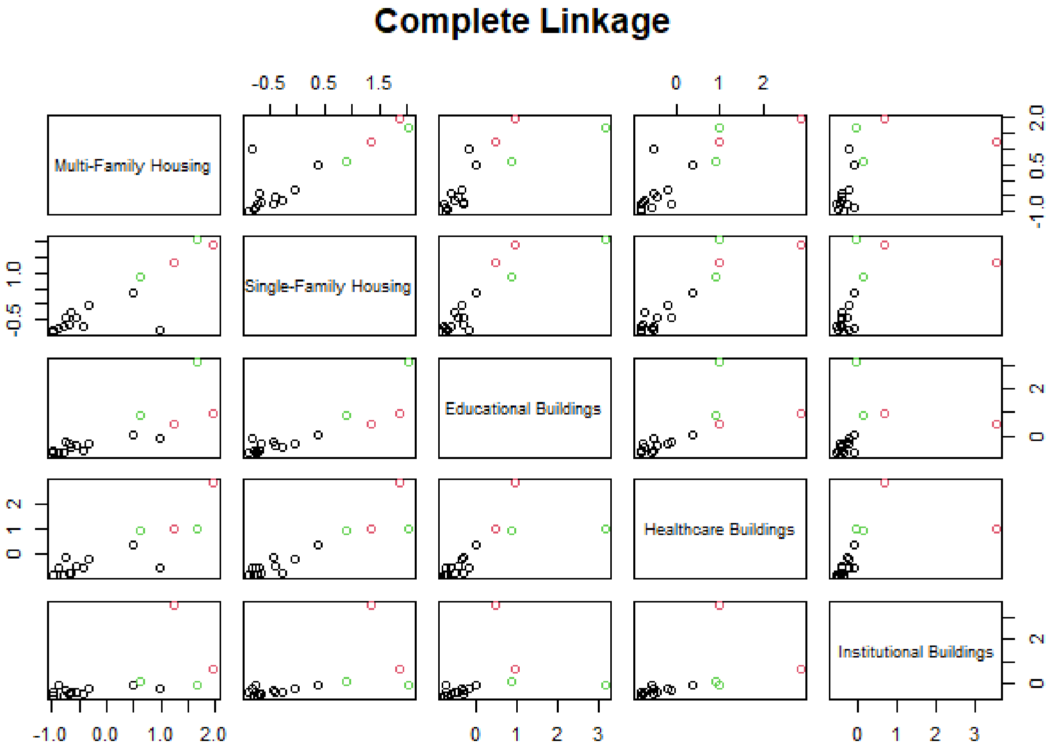

Figure 6: Example of HC: Scatterplot 1, variables 1 to 5

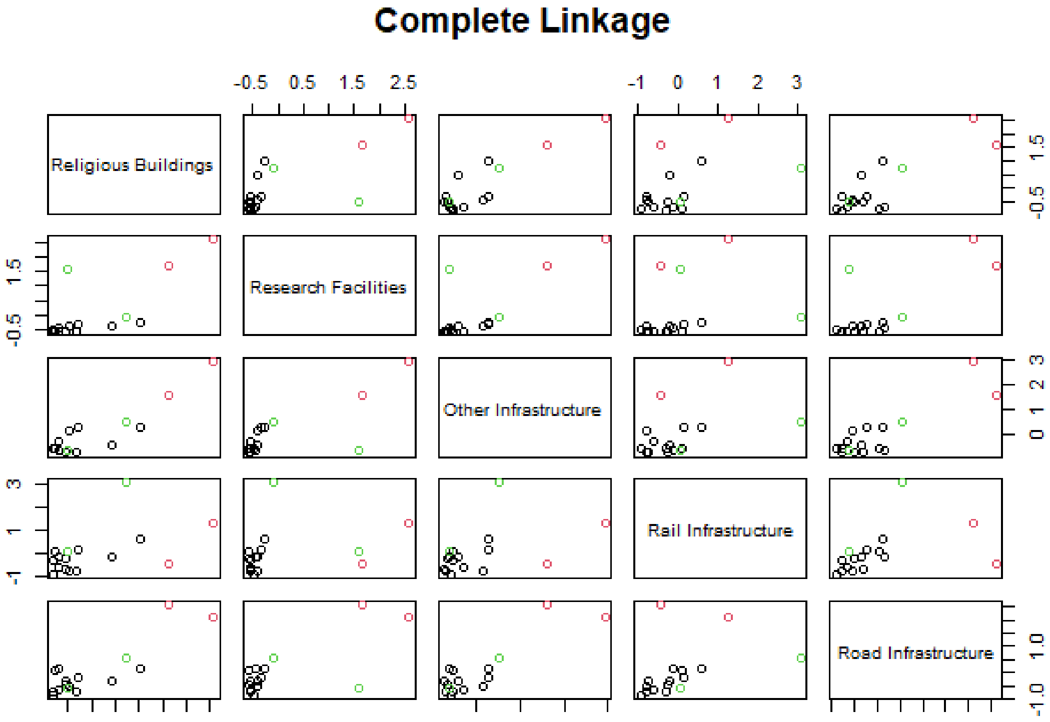

Figure 7: Example of HC: Scatterplot 2, variables 5 to 10

K-means Clustering

For K-mean Clustering, the parameter k is chosen based on the computation of the within variance for values of k ranging from 1 to 10. It is important to note that if k is very high the algorithm will overfit the data, specially in this case where there are only 15 datapoints. Results are shown in Figure 8, 3-mean and 5-mean clustering will be tested upon the data.

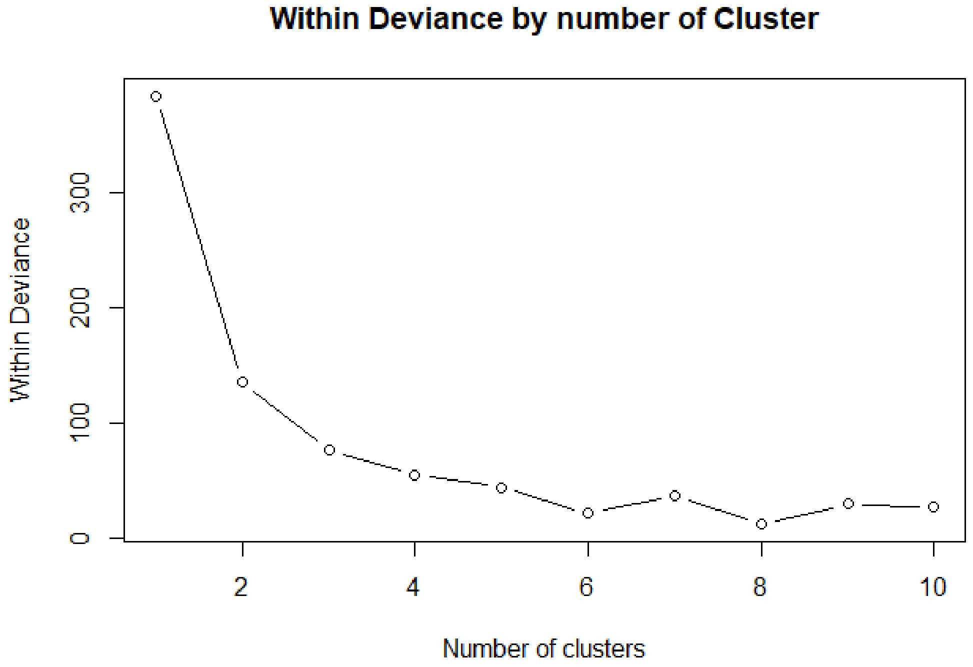

Figure 8: Within Deviance by number of K-clusters

For k=3, a comparison between this classification and the Hierarchical Clustering methods used before can be shown using a confusion matrix. Results in Tables 1, 2 and 3 show virtually the same classification as the Complete Linkage method, a slight difference (4 miss classifications) against Average Linkage and a higher difference against Ward Distance (7 miss classifications).

Table 1: Confusion matrix: Complete Linkage vs 3-mean Clustering

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	12	0	0
Cluster 2	0	0	2
Cluster 3	0	2	0

Table 2: Confusion matrix: Average Linkage vs 3-mean Clustering

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	12	1	0
Cluster 2	0	0	2
Cluster 3	0	1	0

Table 3: Confusion matrix: Ward Distance vs 3-mean Clustering

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	9	0	0
Cluster 2	0	0	2
Cluster 3	3	2	0

For k=5 the confusion matrix analysis is not suitable, as the number of clusters are different than the other techniques. In this case, is better to compare the different methods using a single measurement, the R^2. Results are displayes in Table 4 below.

Table 4: R^2 for Hierarchical Clustering and K-mean Clustering

Method	R^2
Complete Linkage	0.34
Average Linkage	0.39
Ward Distance	0.18
3-mean Clustering	0.68
5-mean Clustering	0.20

3-mean clustering is showing the highest R^2 amongst the tested algorithms, suggesting the more accurate classification of our data. The countries would be assigned the same way they were assigned with the Complete Linkage method. As far as this Unsupervised Learning sections goes, it can be concluded that the behaviour of the equipment market in the construction sector in Turkey is similar in certain end sectors to other European countries, mainly the Netherlands, Italy and Switzerland, therefore suggesting a reference point when conducting country analysis and estimating results.

A final remark must be made. This doesn't mean that the market size of these countries is similar - Turkey has almost x10 the population of Switzerland, to give an example. It means that those markets share some characteristics and can be grouped by them.

Supervised Learning

Given the structure of the database used in this second phase of the project - a 15x263 matrix after data cleansing- the number of algorithms that can be used is limited, either because the datapoints are too few or because the number of features is too large, or because the relation between both is inadequate. In this scenario, simple linear models tend to produce accurate estimates, but some regularization must be applied, as analyzing the features one by one would be a never ending task. Log transformation was used on the dependent variable as well as the most skewed features. All features were scaled to have mean 0 and a standard deviation of 1

Before implementing any algorithm, the 15 datapoints were split into 15 different test folds, leaving in each case a training fold of 14 observations and a test fold of a single observation. These are the folds from which tuning parameters lambda and alpha are chosen and SSE estimated. LOOCV is again used on each training fold to determine the best parameter of each corresponding split.

Lasso Regression

In Table 5 below is displayed the lambda obtained by LOOCV which minimized the MSE (Mean Squared Error), the corresponding prediction using the test fold and the Squared Residual (SR) associated to the prediction

Table 5: Lasso regression results

Fold	Lambda	Prediction	SSE
1	0.14	7.88	0.11
2	0.23	6.50	0.00
3	0.14	7.65	0.16
4	0.12	7.88	1.36
5	0.02	6.49	0.34
6	0.05	8.87	1.11
7	0.11	6.96	0.31
8	0.01	7.78	0.45
9	0.05	8.19	1.35
10	0.31	8.58	0.86
11	0.30	6.98	0.71
12	0.32	8.90	0.78
13	0.12	8.12	0.01
14	0.26	8.46	1.32
15	0.24	9.44	0.00

Average lambda is 0.16 and SSR is 8.8. With this value the final Lasso model is fitted to the whole dataset, the intercept and 10 coefficients are chosen, their value is displayed on Table 6.

Table 6: Lasso coeficients

Feature	Coefficient
(Intercept)	7.92
Gross household savings˜Value	0.06
GVA financial intermediation, real estate and business activities	0.19
Female infant mortality rate per 1000 live births	0.29
Households in income bracket, 75000+ PPP Dollars	0.27
Male population aged 25-29 years	0.04
Male population aged 70-74 years	0.05
Male population aged 75-79 years	0.22
Population density	0.06
Population in income bracket, 20000+ PPP Dollars	0.15
Proportion of male population aged 35-39 years	-0.01

Ridge Regression

For the Ridge regression, the lambdas that minimize MSE obtained by LOOCV are displayed in Table 7. Average lambda is 11.2 and SSR is 5.23. Ridge regression coefficients are not displayed in this report to maintain simplicity.

Table 7: Ridge regression results

Fold	Lambda	Prediction	SE
1	11.66	7.42	0.02
2	11.04	6.74	0.04
3	11.69	7.48	0.33
4	11.43	7.35	0.40
5	11.50	6.93	0.02
6	10.13	9.75	0.03
7	11.69	7.02	0.25
8	11.51	7.47	0.13
9	11.40	7.13	0.01
10	10.78	8.49	1.04
11	10.71	7.02	0.78
12	10.34	9.06	0.52
13	11.62	8.01	0.00
14	11.97	8.60	1.67
15	10.69	9.47	0.00

Elastic Net Regression

Finally, Elastic Net regression is applied to the data. Frist, LOOCV is applied just as before but with 10 different values of alpha (from 0.1 to 1). This gives a 15x10 matrix which is then averaged for each alpha to determine the lambda to be used. Then, LOOCV is applied once again on each fold to determine the alpha with the lowest SSR. Table 8 shows the results: the parameters that minimize the SSR are alpha=0.1 and lambda=0.65

Table 8: Elastic Net Results

alpha	Lambda	SSE
0.1	0.65	4.09
0.2	0.36	4.27
0.3	0.30	4.75
0.4	0.24	5.09
0.5	0.20	5.26
0.6	0.16	5.43
0.7	0.15	5.63
0.8	0.14	6.10
0.9	0.15	6.36
1	0.16	6.83

From the three estimated models, Elastic Net is showing the best results, with an SSR of 4.09 against 8.88 of the Lasso and 5.23 of the Ridge regression. If the R^2 is computed, the result would be (Table 9):

Table 9: R^2 of regularization models

Model	R^2
Lasso	0.59
Ridge	0.76
Elastic Net, a=0.1	0.81

Bootstrap Regression

Even though the best model has been chosen, as the size of the dataset is small, the standard errors of the coefficients are not reliable. For this reason, Bootstrap regression is applied to the Elastic Net model to have an accurate estimation of the standard errors of the coefficients and determine its significance.

The process is executed as follows (Fox, 2015):

1. Predict regression coefficients for the original sample, calculate fitted values and residuals using the Elastic Net model already estimated
2. Bootstrap the residuals 10,000 times and sum each vector to the fitted value of y obtained in the first step
3. Regress the bootstrapped fitted values of y using the Elastic Net regression (optimal lambda and alpha already computed) coefficients against the fixed X matrix of features
4. Compute standard errors and t-statistics to determine significance of the selected features

On Table 11 the following results are shown: Coefficients before applying Bootstrap regression, Coefficients after applying Bootstrap regression, standard errors and t statistics of the Bootstrap regression and if the coefficient is significant at a 95% confidence level for the original coefficients and the Bootstrapped coefficients.

Table 10: Bootstrap Regression of Elastic Net model

Features	Coef. w/o Bootstrap	Coef. w/ Bootstrap	Standard errors	t-statistic	Significant Bootstrap	Significant Elastic Net
(Intercept)	7.97	7.94	0.07	112.99	Yes	Yes
Gross disposable income per household	0.02	0.00	0.01	0.39	No	No
Gross household disposable income	0.01	0.01	0.01	1.61	No	No
Gross household income	0.01	0.02	0.01	2.27	Yes	No
Gross household savings	0.05	0.02	0.02	1.18	No	Yes
Gross national income	0.02	0.02	0.01	2.75	Yes	Yes
Gross national savings	0.01	0.02	0.02	1.09	No	No
National disposable income in current prices	0.02	0.02	0.01	2.88	Yes	Yes
Net national income	0.01	0.02	0.01	2.26	Yes	No
Prop of gross household savings in nominal GDP	0.00	0.00	0.01	0.27	No	No
Nominal GDP	0.02	0.02	0.01	2.95	Yes	Yes
Final consumption expenditure	0.02	0.02	0.01	2.23	Yes	Yes
Final government consumption expenditure	0.01	0.02	0.01	1.64	No	No
Final household consumption expenditure	0.02	0.02	0.01	1.73	No	No
Gross fixed capital formation	0.01	0.02	0.01	1.90	No	No
Prop of gross fixed capital formation in GDP	-0.01	-0.00	0.01	-0.39	No	No
Financial intermediation, real estate and others	0.04	0.03	0.01	2.64	Yes	Yes
Industry	0.00	0.01	0.01	1.23	No	No
Mining, manufacturing and utilities	0.01	0.01	0.01	1.13	No	No
Other services	0.01	0.01	0.01	0.89	No	No
Services	0.02	0.02	0.01	2.94	Yes	Yes
Transport, storage and communications	0.01	0.03	0.02	1.74	No	No
Total GVA	0.02	0.02	0.01	2.72	Yes	Yes
Agriculture, forestry and fishing in GVA	-0.07	-0.03	0.03	-0.93	No	No
Financial intermediation in GVA	0.04	0.03	0.03	0.96	No	No
Real GDP	0.02	0.02	0.01	2.12	Yes	Yes
Real household consumption	0.02	0.01	0.01	1.53	No	Yes
Female births	0.00	0.02	0.01	1.41	No	No
Female infant mortality rate	0.15	0.11	0.06	1.94	No	Yes
Female pop	0.01	0.01	0.01	1.56	No	No
Female pop aged 25-29 years	0.01	0.02	0.01	1.44	No	No
Female pop aged 30-34 years	0.01	0.01	0.01	1.39	No	No
Female pop aged 50-54 years	0.01	0.00	0.00	1.12	No	Yes
Female pop aged 55-59 years	0.01	0.01	0.01	1.33	No	No
Female pop aged 60-64 years	0.01	0.01	0.00	1.27	No	No
Female pop aged 65-69 years	0.01	0.01	0.01	1.50	No	No
Female pop aged 70-74 years	0.02	0.01	0.01	1.74	No	Yes
Female pop aged 75-79 years	0.01	0.01	0.01	1.24	No	Yes
Female pop aged 85-89 years	0.00	0.00	0.00	0.64	No	No
Female pop aged 90-94 years	0.00	0.00	0.01	0.63	No	No
Households in income bracket, 40000-75000	0.01	0.02	0.01	1.29	No	No
Households in income bracket, 75000+	0.13	0.06	0.03	2.10	Yes	Yes
Infant mortality rate	0.14	0.10	0.05	2.20	Yes	Yes
Life expectancy at birth, female	-0.12	-0.04	0.06	-0.70	No	No
Male births	0.00	0.02	0.01	1.45	No	No
Male infant mortality rate	0.12	0.08	0.05	1.66	No	Yes
Male pop	0.01	0.01	0.01	1.68	No	No

Table 11: Bootstrap Regression of Elastic Net model

Features	Coef. w/o Bootstrap	Coef. w/ Bootstrap	Standard errors	t-statistic	Significant Bootstrap	Significant Elastic Net
Male pop aged 25-29 years	0.02	0.02	0.01	1.61	No	No
Male pop aged 30-34 years	0.01	0.01	0.01	1.50	No	No
Male pop aged 50-54 years	0.00	0.00	0.00	0.91	No	No
Male pop aged 55-59 years	0.01	0.01	0.01	1.19	No	No
Male pop aged 60-64 years	0.01	0.01	0.01	1.27	No	No
Male pop aged 65-69 years	0.01	0.01	0.01	1.90	No	Yes
Male pop aged 70-74 years	0.03	0.02	0.01	2.03	Yes	Yes
Male pop aged 75-79 years	0.02	0.01	0.01	1.62	No	Yes
Male pop aged 85-89 years	0.01	0.00	0.00	0.91	No	Yes
Male pop aged 90-94 years	0.00	0.00	0.01	0.66	No	No
Number of households	0.01	0.01	0.01	1.80	No	No
Number of households with children	0.05	0.02	0.02	1.11	No	Yes
Pop	0.01	0.01	0.01	1.63	No	No
Pop aged 15-64 years	0.00	0.01	0.00	1.34	No	No
Pop aged 25-29 years	0.01	0.02	0.01	1.54	No	No
Pop aged 30-34 years	0.01	0.01	0.01	1.48	No	No
Pop aged 50-54 years	0.01	0.00	0.00	1.03	No	No
Pop aged 55-59 years	0.01	0.01	0.01	1.30	No	No
Pop aged 60-64 years	0.01	0.01	0.00	1.33	No	No
Pop aged 65+ years	0.02	0.01	0.01	1.81	No	Yes
Pop aged 65-69 years	0.01	0.01	0.01	1.74	No	Yes
Pop aged 70-74 years	0.02	0.01	0.01	1.91	No	Yes
Pop aged 75-79 years	0.01	0.01	0.01	1.46	No	Yes
Pop aged 85-89 years	0.01	0.00	0.00	0.74	No	No
Pop aged 90-94 years	0.00	0.00	0.01	0.66	No	No
Pop density	0.02	0.03	0.02	1.23	No	No
Pop in income bracket, 20000+ PPP Dollars	0.05	0.05	0.02	2.63	Yes	Yes
Prop of female population aged 35-39 years	-0.01	-0.00	0.01	-0.42	No	No
Prop of households, 4000-10000 PPP Dollars	-0.05	-0.03	0.03	-1.04	No	No
Prop of male pop aged 35-39 years	-0.07	-0.03	0.03	-1.12	No	No
Prop of male pop aged 40-44 years	-0.02	-0.02	0.02	-1.13	No	No
Prop of pop aged 35-39 years	-0.04	-0.01	0.01	-0.73	No	No
Prop of pop aged 40-44 years	-0.00	-0.01	0.02	-0.85	No	No
Rural pop	0.01	0.01	0.01	0.62	No	No
Total births	0.00	0.02	0.01	1.44	No	No
Total deaths	0.01	0.01	0.01	1.10	No	No
Urban pop	0.00	0.01	0.01	1.32	No	No

Now that the standard errors have been properly estimated, a more robust conclusion can be drawn about the coefficients that are really significant for the Elastic Net model estimated.

By this point of the project the estimation of the market size of Mini Excavators in Turkey can be made. Model result is 9.7, which when reversed the log transformation gives 16,332.

Conclusions and further steps

Using Statistical Learning Techniques two goals have been accomplished: the first one, find similarities among construction markets of different countries and second, estimate a market size for which there is no another estimate available. For sure there is room for improvement in both cases, beginning with the data selection and data pre processing, maybe a different approach would yield more robust results.

Regarding the techniques used, it is a clear limitation the size of the dataset, but this can be overcome by applying well tailored techniques with tuned parameters. Further linear models can be explored and even non parametric models could produce a good estimation of the market.

Finally, the output of the project is only as valid as it is useful for businesses. With this in mind, an opportunity to improve would be to find a less computationally intensive workflow and a model that's easier to interpret. The number estimated here must also be confronted with experts on the field to get their feedback and declare it reliable. No market size estimation is exact, but it must be close with certain degree to be considered truthful

Bibliography

References

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.