A group of men playing a game of football

Description generated with high confidence

**Analysis on Soccer Player Attributes**

**By: Sau Kha**

**April 2018**



# Executive Summary

The research questions are:

Fifteen blocks of huge data from smart meters of 5,567 London households were explored, analyzed and visualized. Python programming tools, mainly the Pandas and Matplotlib modules, were used.

Appendix A presents all the Python scripts used in this study. Appendix B displays charts generated from the dataset.

# Table of Contents

[Executive Summary i](#_Toc512109105)

[Introduction 1](#_Toc512109106)

[Research Questions 1](#_Toc512109107)

[Reference 1](#_Toc512109108)

[Methodology and Analysis Approach 2](#_Toc512109109)

[Findings 4](#_Toc512109110)

[Re-Grouping of Acorn Types in Broader Categories 4](#_Toc512109111)

[Answer to Research Question 1 4](#_Toc512109112)

[Answer to Research Question 2 4](#_Toc512109113)

[Answer to Research Question 3 4](#_Toc512109114)

[Conclusion 4](#_Toc512109115)

[APPENDIX A: Python Scripts 5](#_Toc512109116)

[APPENDIX B: Source of Data 7](#_Toc512109117)

List of Charts and Tables

[Table 1. Distribution of Smart Meter Users in Acorn Category 4](#_Toc501101406)

[Chart A. Total Electricity Usage by Acorn Category 5](#_Toc501101407)

[Table 2. Overall Usage Distribution by Acorn Category 5](#_Toc501101408)

[Chart B. Total Electricity Usage by Acorn Type 6](#_Toc501101409)

[Table 3. Overall Usage Distribution by Acorn Type 6](#_Toc501101410)

List of Figures

Figure 1: Total Electricity Usage by Acorn Category – Block\_0-a 19

Figure 2: Total Electricity Usage by Acorn Category – Block\_0-b 19

Figure 3: Total Electricity Usage by Acorn Category – Block\_1 20

Figure 4: Total Electricity Usage by Acorn Category – Block\_2 20

Figure 5: Total Electricity Usage by Acorn Category – Block\_3 21

Figure 6: Total Electricity Usage by Acorn Category – Block\_4 21

Figure 7: Total Electricity Usage by Acorn Category – Block\_5 22

Figure 8: Total Electricity Usage by Acorn Category – Block\_6 22

Figure 9: Total Electricity Usage by Acorn Category – Block\_7 23

Figure 10: Total Electricity Usage by Acorn Category – Block\_8 23

Figure 11: Total Electricity Usage by Acorn Category – Block\_9 24

Figure 12: Total Electricity Usage by Acorn Category – Block\_10 24

Figure 13: Total Electricity Usage by Acorn Category – Block\_11 25

Figure 14: Total Electricity Usage by Acorn Category – Block\_12 25

Figure 15: Total Electricity Usage by Acorn Category – Block\_13 26

Figure 16: Total Electricity Usage by Acorn Category – Block\_14 26

Figure 17: Total Electricity Usage by Acorn Category by Month – Block\_0-a 27

Figure 18: Total Electricity Usage by Acorn Category by Month – Block\_0-b 27

Figure 19: Total Electricity Usage by Acorn Category by Month – Block\_1 28

Figure 20: Total Electricity Usage by Acorn Category by Month – Block\_2 28

Figure 21: Total Electricity Usage by Acorn Category by Month – Block\_3 29

Figure 22: Total Electricity Usage by Acorn Category by Month – Block\_4 29

Figure 23: Total Electricity Usage by Acorn Category by Month – Block\_5 30

Figure 24: Total Electricity Usage by Acorn Category by Month – Block\_6 30

Figure 25: Total Electricity Usage by Acorn Category by Month – Block\_7 31

Figure 26: Total Electricity Usage by Acorn Category by Month – Block\_8 31

Figure 27: Total Electricity Usage by Acorn Category by Month – Block\_9 32

Figure 28: Total Electricity Usage by Acorn Category by Month – Block\_10 32

Figure 29: Total Electricity Usage by Acorn Category by Month – Block\_11 33

Figure 30: Total Electricity Usage by Acorn Category by Month – Block\_12 33

Figure 31: Total Electricity Usage by Acorn Category by Month – Block\_13 34

Figure 32: Total Electricity Usage by Acorn Category by Month – Block\_14 34

Figure 33: Total Electricity Usage by Acorn Type – Block\_0-a 35

Figure 34: Total Electricity Usage by Acorn Type – Block\_0-b 35

Figure 35: Total Electricity Usage by Acorn Type – Block\_1 36

Figure 36: Total Electricity Usage by Acorn Type – Block\_2 36

Figure 37: Total Electricity Usage by Acorn Type – Block\_3 37

Figure 38: Total Electricity Usage by Acorn Type – Block\_4 37

Figure 39: Total Electricity Usage by Acorn Type – Block\_5 38

Figure 40: Total Electricity Usage by Acorn Type – Block\_6 38

Figure 41: Total Electricity Usage by Acorn Type – Block\_7 39

Figure 42: Total Electricity Usage by Acorn Type – Block\_8 39

Figure 43: Total Electricity Usage by Acorn Type – Block\_9 40

Figure 44: Total Electricity Usage by Acorn Type – Block\_10 40

Figure 45: Total Electricity Usage by Acorn Type – Block\_11 41

Figure 46: Total Electricity Usage by Acorn Type – Block\_12 41

Figure 47: Total Electricity Usage by Acorn Type – Block\_13 42

Figure 48: Total Electricity Usage by Acorn Type – Block\_14 42

# Introduction

This purpose of this project is to study the attributes of soccer players of an European soccer database from Kaggle.com.

Research questions

This document reports findings from a sample of 5,567 London households that took part in a pilot project between November 2011 and February 2014.

Smart meter data from participant households and their household classification based on an ACORN approach [Reference 1 and 2] are analyzed. Findings in relation to amount and pattern of usage may be used to strategically plan future energy development.

## Research Questions

Research questions for this study include:

1. According to the Acorn classification, which category of household is the biggest user of electricity for the time frame of the study?
2. Within the category, which type of household is the biggest user of electricity?
3. Which are the peak usage months?

After exploring the collected data, plans were to acquire additional data, such as weather and life styles of household to explore possible explanations for the observations.

Intended audience of this study include government policy makers, energy providers, energy conservation groups and electricity consumers. Findings from the study may be used to make strategic energy development plan for different areas of the country and consumer-targeted promotions of energy savers consumer plans, energy efficient cars and appliances, energy generators, solar panels, and so on. This type of study should be conducted in a continuous basis, so that strategic plans may be adjusted as the situation changes.

# Reference

1. “The Acorn User Guide, The Consumer Classification”, CACI Ltd, 2014.
2. “Acorn Technical Guide”, CACI Ltd, 3/2/2017.

# Methodology and Analysis Approach

The methodology for this study basically is first to explore, clean up, analyze, visualize the data and then draw conclusions and attempt to find answers to the research questions. The following steps were taken to reach the findings and conclusions:

1. Unzip the data files and write Python scripts to explore and clean up data. To progressively approach the study, data files were written over with cleaned up file versions. These include all 15 blocks of smart meter data and the information household filed.
   1. Data Clean Up: Any null, blank and erroneous record are dropped. This include user data with incomplete Acorn Types, such as ACORN- and ACORN-U, that do not relate to any Acorn Category.

Lesson learned: This was done because most of the time they cause error when a Python function or method is attempted.

* 1. Information Extraction and Exploration: The timestamp column from the text smart reader files were converted to datetime datatype. As each data file was explored and cleaned up, I used Pandas function (pd.datetime) to extract Year, Month and Hour information from the datetime column. Year, Month and Hour information were saved as added columns in the csv file for later data analysis.

Lesson Learned: The additional columns were not necessary, as the same information could be obtained during data analysis phase. This also have made the source text files bigger and cost longer or memory problem during analysis. For example, the first zip file of 15 blocks of smart meter data, block\_0.csv.gz, has the largest file size. After extracting the file and adding extra columns, the resulting text file when read as a Pandas DataFrame, block\_0.csv was one of the few that consistently caused my computer to freeze. The IPython kernel, the Python execution backend for Jupyter notebook frequently died or automatically restarted.

* 1. Data Merging and Analysis: Household information of participants were merged with the smart meter data, so that data could be analyzed per Acorn Type and Category. Lesson Learned:
     1. While it was possible read and analyze data with Pandas, it would be much better if a relational database was created to import data from each block as a table in the database. Together with the household information imported as a separate table, queries could be done to pull data from different tables for better data analysis. SQLite3 is a very good tool for this type of analysis.
     2. With memory limitation, I considered using the pickle module in Python. The idea was to convert a slice of information, such as the total electricity usage for a certain time frame for a certain Acorn type or category, from a block of smart meter data into Pandas DataFrame and then to save it as pickle file. After the same slice of information was pulled from each block and saved as a pickle file, all pickle could be concatenated and visualized.

The [pickle](https://docs.python.org/3/library/pickle.html#module-pickle) module implements binary protocols for serializing and de-serializing a Python object structure. “Pickling” is the process whereby a Python object hierarchy is converted into a byte stream, and “unpickling” is the inverse operation, whereby a byte stream (from a [binary file](https://docs.python.org/3/glossary.html#term-binary-file) or [bytes-like object](https://docs.python.org/3/glossary.html#term-bytes-like-object)) is converted back into an object hierarchy. Pickling (and unpickling) is alternatively known as “serialization”, “marshalling,” or “flattening”; however, to avoid confusion, the terms used here are “pickling” and “unpickling”. *Source: Python Standard Library 12.1 Python Object Serialization”* [***https://docs.python.org/3/library/pickle.html***](https://docs.python.org/3/library/pickle.html)

* + 1. With memory problem encountered early in the project, I decided to just analyze each block of smart meter data separately. If each block of data was an unbiased sample of the population, the analysis result from each block should be comparable to each other. Otherwise, further analysis would be needed to obtain overall results from the complete dataset. In the last minutes as I was writing this report, I was able to use Pandas “groupby” mechanism to systematically aggregate summary data and save them in two csv files in Python. That was completed by writing and running scripts directly on Jupyter Notebook. The next set of groupby scripts were run by revising what were on an untitled Jupyter Notebook. This could have been done probably in a well-planned and efficient programming way.
  1. Data Visualization: Charts were created from each block of data and compared among each other using the Matplotlib module.
  2. Draw Conclusion: Find answers to the research questions of this study and draw conclusions.

# Findings

## Re-Grouping of Acorn Types in Broader Categories

## Answer to Research Question 1

## Answer to Research Question 2

## Answer to Research Question 3

# Conclusion

In short, analysis of the dataset provides answers to all three research questions. Affluent consumers are the biggest users of electricity. And within the Affluent category, Acorn Type E, Career Climbers, use the most energy. Additional information concerning characteristics and lifestyles of Career Climbers may be obtained from the Acorn User Guide (Reference 1).

People’s career opportunity, family situation, financial obligations and living area and housing styles could very well explain the high energy usage of this group of people. Besides, participants of the Smart Meter pilot study did not acquire the meter at the same time. So, comparing the total usage from among Acorn categories or types may not be comparing apples to apples. Participants who started in the winter months would have logged more usage in the meter than those who started in the summer months the following year. Likewise, those who participated in the early phase of the pilot study will have a greater total usage. Use of more sophisticated programming tools will be able to assist with more precise data analysis. The total number of individuals in Acorn categories and types will certainly affect the total usage for the group. It is better to the calculate the average usage by dividing the total usage by the number of participating household. The number of family members in the household, the region and the type of housing, and so on, may have an impact on energy usage. However, this type of information is not available for the study.

As far as the peak usage in relation the weather in certain months of the year, it was pretty much predicted. Future work to continue this study includes investigating the use of API to collect weather data.

The major weakness of the study is the lack of proficiency in Python packages, such as sqlite3 or SQLAlchemy. These tools will allow queries to be run to gather relational information from different tables. In the case of the current study, information from different data blocks, may be assembled according to Acorn Category, Type, year, month or time of day and analyzed.

# APPENDIX A: Python Scripts

# APPENDIX B: Source of Data

Data for this project was downloaded from this website in SQLite3 format: <https://www.kaggle.com/hugomathien/soccer>

The following presents brief information of the database:

* +25,000 matches
* +10,000 players
* 11 European Countries with their lead championship
* Seasons 2008 to 2016
* Players and Teams' attributes sourced from EA Sports' FIFA video game series #
* Team line up with squad formation (X, Y coordinates)
* Betting odds from up to 10 providers
* Detailed match events (goal types, possession, corner, cross, fouls, cards etc...) for +10,000 matches

\**16th Oct 2016: New table containing teams' attributes from FIFA !*

# Original data source for players and teams attributes tables: <http://sofifa.com/> : players and teams attributes from EA Sports FIFA games. FIFA series and all FIFA assets property of EA Sports. Foreign keys "api\_id" for players and matches are the same as the original data sources.