Aston University
Birmingham B4 7ET
http://aston.ac.uk/

Exploratory analysis challenge

KTP XYZ

Technical Task July 11, 2023

Introduction

The goal of this challenge is to explore a complex data set related to an industrial application. You will receive a set of data tables in the form of .xlsx files, containing different information about the problem.

This challenge must be completed **individually**. You can use Python or R for your development. Your analysis must be fully and easily reproducible from the data files provided, which may be best achieved by using a literate programming format to communicate your solution, e.g., a Jupyter Notebook or R Markdown document.

Task Details

Your task is to explore a set of relational data tables related to failure events in a fleet of high-power electrical transformers from a South American distribution utility, to investigate which factors may be associated with an increased probability of two failure modes:

- **Early failures**, which occur in the first three years after an equipment is commissioned into operation. These failures are often (although not always) associated with constructive problems.
- **Wear-and-tear failures**, which occur after at least three years of operation, and are commonly associated to a combination of the usual wear-and-tear due to electrical stresses experienced by the device and possible cumulative suboptimal maintenance.

The data is composed of eight Excel worksheets. Seven of these relate to events that happened in different operations areas. These files have names in the format *Events_XY.xlsx*, and each contains 5 columns:

1. **Area Code**, which contains the code of the operational area in which the events detailed in that table occurred.

2. **Op ID**, a unique ID of each event.
3. **Op Date**, the date in which the event was registered.
4. **Equipment ID**, the ID of the equipment that generated the event.
5. **Fail Flag**, which indicates whether the event registered was a destructive failure.

The remaining file, *Devices.xlsx*, contains the specs and some operational information on all devices. This file has the following columns:

1. **ID** (string), the unique ID of each equipment
2. **TYPE** (string), the type of the equipment
3. **AREA** (string), the area where the equipment is installed
4. **FUNCTION** (string), the primary purpose of the equipment
5. **MANUFACTURER** (string), the (anonymised) manufacturer of the equipment
6. **INSTALL DATE** (date), the date on which the equipment started operating.
7. **PHASES** (string), the number of phases of the equipment.
8. **POWER (MVA)** (numeric), the power rating of the device (in million-Volt-Amperes)
9. **VOLTAGE REGULATION** (binary), an indicator variable informing whether the device has voltage regulation
10. **NOMINAL VOLTAGE (KV)** (numeric), voltage rating of the device (in thousand-Volts)

The main technical goals in this challenge are (i) to explore which factors (e.g., AREA, FUNCTION, POWER, MANUFACTURER, etc.), if any, are strongly associated with each failure mode (early failures / wear-and-tear failures); (ii) to discuss which variable values could represent good predictors of failure (for each mode); and (iii) to suggest evidence-based recommendations. More specifically, your solution must contain:

1. A description of your preliminary exploration of the data, including identification of data inconsistencies (e.g., failure dates earlier than the corresponding install dates, aliasing in categorical variables, duplicated entries, etc.), missing data, and any other problem that may be detected in the data.
2. A description of your data preprocessing - how you remedied the data inconsistencies identified in the preliminary exploration, what was your policy for missing values, etc.
3. A qualitative / descriptive analysis of your data, in which you will employ graphics and summary statistics to investigate some possible hypotheses you may have regarding the main factors associated with the two failure modes of interest here, namely *early failures* and *wear-and-tear failures.*

4. Some type of modelling of your data. This could be, for instance, some type of regression, survival analysis (trying to estimate the time-to-failure of individual devices), clustering (trying to identify groups of devices that share common characteristics, and possibly correlate these groups with the amount of observed failures), etc.. There is no specific modelling type required here - you are only required to perform *some* modelling using what you have learned about the data in your early exploration. Your choice of modelling should be justified in your solution (i.e., the model must be related to some sort of question you may want to investigate in the data).

5. Any recommendations based on patterns you have discovered in the data, as well as some reflection on the strengths and weaknesses of your own analysis.

Although not directly related to the project you will be working on, the data in this challenge represent a good example of the challenges of working with data collected by multiple platforms over long periods of time.

Tips and analysis pitfalls

Some of the important details and pitfalls to be aware of are:

- There is missing data in several variables. You can choose to treat missing data by listwise deletion, imputation, variable deletion or any other strategy you consider relevant. Please justify your choice in your solution.
- There may be duplicated or inconsistent entries (e.g., devices with two distinct events flagged as a destructive failure, non-unique ID entries in the *Devices.xlsx* table, etc.). As with missing data, you can choose how to deal with those, and your decision should be documented and justified.
- Dates are expressed in a variety of inconsistent formats. Dealing with this issue is part of the challenge.
- Manufacturer names are expressed inconsistently (e.g., “Manufacturer01”, “manuf-1” and “manufacturer 01” all refer to the same manufacturer). Dealing with this issue is part of the challenge. You may want to explore all unique values in the manufacturer variable, define an alias table and use it to consolidate manufacturer names.
- Notice that commas are used as decimal separators (instead of points). It is important to consider and correct this when importing, exploring and preprocessing your data.
- Failure flags in the *Events* tables are indicated either by the presence of an X (TRUE) or a missing value (which should be interpreted as FALSE).
- Notice that the same equipment can appear multiple times in a given *Events* table - it may have generated several events before it finally failed (if it did). Notice, however, that only a single event indicated as a **failure** (i.e., having an ”X” in variable **Fail Flag**) can exist for any individual equipment. Duplicated failures should be treated as data inconsistencies.

A few further suggestions that may be useful as you complete this challenge:

- Carry out your analysis in a systematic way. Take good notes on what you have done and the rationale behind your choices, so that you can fully document it later.
- Read the information provided in this document carefully, and keep it in mind as you explore the data.
- Be sure to spend some time exploring and understanding the dataset, and consider carefully how you will treat outliers, missing data, or any other important aspects of this data.
- Remember that you are interested in investigating *two* distinct problems: the *early failures*, i.e., failures that occur less than three years after a device is commissioned into service; and *wear and tear failures*, which are those that occur after more than three years of operation.

1 Submission

The submission should be a reproducible report - either a notebook-type analysis document (+ any additional scripts developed as part of your solution) or a pdf document accompanied by a set of analysis scripts, with instructions on how to reproduce your analysis. In summary, you are expected to submit:

- Your **report**, which can be in one of the following formats:
  - Jupyter Notebook (IPYNB)
  - R Markdown document (RMD)
  - PDF document
- Any other files that may be required for the full replication of your analysis and results. This can include data consolidation and cleanup scripts, a data table of aliases for manufacturer names, or any other file required for us to fully replicate your analysis process.

Notice that regardless of the report format, it should be a coherent, consistent report (not only a collection of code chunks). Generally speaking, your report should contain the following sections:

Executive Summary A brief description (2 or 3 paragraphs) of the problem being investigated, the general approach followed, and the main results of your analysis.

Problem Definition A general overview of the problem being investigated.

Exploratory Data Analysis A description of your preliminary investigation of the data: what you did and what you found out regarding, e.g., missing data, data inconsistencies, number of observations, number of failure events of each type, etc.)

Data Preprocessing A description of your data preprocessing steps, including a justification for your analysis decisions regarding, e.g., the treatment of missing values and duplicated entries, the treatment of data inconsistencies, the standardisation of dates, the consolidation of aliased categorical variables, etc.

Descriptive Analysis A preliminary (possibly qualitative) investigation of the information contained in your preprocessed data, including graphical investigation of factors associated with each type of failure, of the the main predictors of failure, etc.

Modelling The definition of the hypothesis you are interested in investigating by modelling, a brief description of the chosen model, the training of the model, and the results obtained. There is no specific “right” model here - you are only asked to select a model that is reasonably adequate for the question you want to investigate (and that this is justified in your report - see *Task Details*). Another **suggestion** is that you keep your modelling simple - a simple but well-reasoned model is much better than an unnecessarily complex one.

Conclusions and Recommendations What you have learned about the data, and your recommendations based on the evidence you have provided in your analysis. This should also contain some reflection on the strengths and weaknesses of your own analysis.