Assignment 2 - Notes

Find the source of the datasets

# Weather.csv

**Missing Data**

-Multiple values missing under the “9am wind direction” column

-Multiple values missing from last row (176)

# Price\_and\_demand

**Missing Data**

-Headings missing

Link: https://aemo.com.au/en/energy-systems/electricity/national-electricity-market-nem/data-nem/data-dashboard-nem

Original Data: <https://aemo.com.au/energy-systems/electricity/national-electricity-market-nem/data-nem/aggregated-data>

-Down the bottom there's a DOWNLOAD HISTORICAL DATA option. Select Nov 2022 / dates

This shows us the missing headers for our price\_and\_demand.csv are:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| REGION | SETTLEMENTDATE | TOTALDEMAND | RRP | PERIODTYPE |

I believe the TOTALDEMAND is in (MW) but will need to be double checked.

Definition of RRP in the

<https://visualisations.aemo.com.au/aemo/di-help/Content/Data_Model/MMS_Data_Model.htm?TocPath=_____8>

under: <https://visualisations.aemo.com.au/aemo/di-help/Content/Data_Model/MMS_Data_Model_Upgrade_Report_52.pdf>

Don't know exactly what the RRP represents - what is it $ in proportion to?

**RRP = regional reference price?**

Refers to price per MWh (mega watt hour) – called them up

TOTALDEMAND is in MW

**The original data is for every 5 mins, whereas Geela has cleaned it up so we only see every 30 mins in the data file she provided us.**

**-she has not aggregated it, simply just selected each 30min interval, this contributes to inaccuracy**

So.. might be important for our report..   
Instead of doing an average of the 6 x 5min sets in each half hour, Geela has just put the data for each 30 min interval and ignored all the "in betweens" if that makes sense.   
This is important because we can mention this in our report as a source of inaccuracy as we've essentially cherry-picked intervals of time while ignoring the data points inbetween,   
and because the RRP seems to deviate significantly (for 4000 MW of demand the RRP could be either -38 or 82 for example).

**AEMO Data Model Reports**

<https://visualisations.aemo.com.au/aemo/di-help/Content/Data_Model/MMS_Data_Model.htm?TocPath=_____8>

PEARSON’S CORRELATION  
  
HIGHEST

3PM TEMPERATURE

# 03 - Data Cleaning

Notes from last night and Lecture on Data Cleaning

03 – Data Cleaning

Goal: Go from raw data to dataframe

Step 1 – determine how the data appears at a glance

**Accuracy** - correct, wrong, accurate or not

* Price\_and\_demand appears to be accurate – can’t verify without expert opinion
* Weather appears to be accurate – can’t verify without expert opinion

**Completeness** - are there missing values?

* Price\_and\_demand appears to be complete – there are a few 0 values but they might be genuine 0 values and not null
* Weather is incomplete and has missing values
  + Some values missing from 9am wind direction and wind speed – specifically rows 91, 103, 104, 113, 115, 119, 132, 147, 149
    - Given basic domain knowledge, “calm” string in wind speed column would likely be 0
    - Given basic domain knowledge, the missing values for wind direction could be imputed
  + Entire column contents are missing – suggest we remove
  + Wind direction at 9am seems to be leanings towards N directions
  + Wind direction at 3pm seems to be leaning heavily towards S directions
  + Numerous data points for Row 176 (the last row) is missing
    - Perhaps we could find the original data source to replace, or omit that row

**Consistency** – date/time consistent? All English characters? Float / str input validation?

* Price\_and\_demand appears to be consistent
* Weather is not consistent
  + For column “wind speed”, there are some strings mixed in – “calm”.
    - Can possibly replace with 0’s

**Timeliness –** is it being updated in a timely manner?

* Price\_and\_demand appears to be updated within 24 hours of release if there are any changes (need source – AEMO)
* Weather – unsure if this has been updated

**Believability** – is it believable or realistic?

* Price\_and\_demand appears to be realistic, although expert knowledge is needed to verify
* Weather appears to be realistic, although expert knowledge is needed to verify (relating to air pressure .etc)

**Interpretability** - how easily can I understand the data?

* Price\_and\_demand seems easy to understand when you find the missing headers
* Weather appears easy to understand

## What type of Missing Data is it?

## Types of Missing Data

**Missing completely at random** - seemingly unrelated to anything

**Missing at random** - likely missing relating to other variables i.e. a lot of males have missing

weight data, but most females have weight data

**Missing not at random** - i.e. as soon as the data goes into triple digits, there is no data .etc

Relating to the actual variable itself

MNAR –

-The “9am wind direction” values appear to be missing directly relating to the column “9am wind speed” – when the wind speed is “calm”, there is no value for wind direction. It’s unclear which value is collected first and therefore which influences the other

-The data from the last row (176) appears to be missing in certain parts – particularly columns

D, H, I, J, Q, R, T, U, V, directly relating to:

D – Maximum temperature

H – Direction of maximum wind gust

I - Speed of maximum wind gust (km/h)

J - Time of maximum wind gust

Q - 3pm Temperature (°C)

R - 3pm relative humidity (%)

T - 3pm wind direction

U - 3pm wind speed (km/h)

V - 3pm MSL pressure (hPa)

Given the non-randomness of the missing data, it seems that the maximum values for the day have not been collected, suggesting that this was very recent data at the time of collection, perhaps the same day

Also given the lack of data from 3pm, this suggests that measurements had not been collected / published for the afternoon, which lends credibility to this presumption

The question is then whether or not to impute the data or to choose not to work with that particular column

PRICE AND DEMAND CLEANING

We want it to be formattable with weather, so we need to convert

**WEATHER DATA**

Need to find missing values

**-Check the original dataset for them**

What other columns should I drop?

# GET STARTED ON DATA DESCRIPTION

Create a Jupyter Notebook file – report

**Go through the lecture first**

Open the lecture that has step-by-step - workshop

# What techniques should we use?

One method is Statistical Measurement Imputation

-Mean

-Median

You can also use TOOLS to clean your data!

Google Refine,

Kettle,

Talend .etc

Go have a look at the last team’s report

GET

Variety of techniques

Outliers - scale, remove or convert them to

Join data together

Raw text to tabular format

No one technique is the one - depends on the situation

Learn more about cleaning on your own

Practice a lot of this with your team members and **Google**

Go from raw data to dataframe

Use Google for data cleaning

Majority of the marks will be placed on step 03 and step 04

What to do first?

Import Data into DF?

Get SKLearn?

Clean Data?

# EOD E-Mail

One of the questions we need to ask is – do we want to impute the data for the last row – April 24th 2023? Given the weather varies a lot, it might be more accurate to simply remove that datapoint, given that most likely the variables we want (max temp .etc) are not available (max temp).

Very likely the RRP col in the price\_and\_demand won’t matter, and that TOTALDEMAND will be the only col that matters, as the previous group’s assingment’s price\_and\_demand.csv didn’t have an RRP col at all