Coursework 1 - Solution and Marking Guidance

Christian Rohrbeck

12 February 2025

```
library(dplyr)
library(tidyr)
library(lubridate)
library(ggplot2)
```

Question 1

Several areas of Utopia are historically suffering from poor air quality. The society "Clean Air for Utopia" has measured PM10 levels (in $\mu g/m^3$) at six sites across Utopia on a daily basis for several years. They now ask you to perform an analysis which explores how the six sites compare in terms of the recorded PM10 levels.

The society is further interested in identifying factors which may explain differences across the sites. They are aware that their current data does not allow them to make such conclusions. You are thus asked to suggest other data or information the society may want to collect so that they can address this question in the future.

Rationale

This question assesses the student's understanding of key aspects covered in the course: reporting the amount of missing data, converting a character to a date object and visualizing the data for several variables in a plot. Students have seen all these aspects in examples in the lecture, but it is expected that a large proportion of students will struggle to present results in a concise manner. The closest example students will have seen is the analysis of weather variables for five Australian cities, but we rarely considered the full data set there and the analysis was quite elaborate. There will also be some students who forget to report the proportion of missing data.

Example of a first class solution

We start by loading the data

```
PM10 <- read.csv( "PM10.csv" )
```

The first step is to report the amount of missing data

```
mean( is.na(PM10$Site1) )
```

```
## [1] 0.03012048
```

```
mean( is.na(PM10$Site2) )
```

```
## [1] 0
```

```
mean( is.na(PM10$Site3) )
```

```
## [1] 0
```

```
mean( is.na(PM10$Site4) )
```

```
## [1] 0.02081051
```

```r
mean( is.na(PM10$Site5) )
```

```
## [1] 0
```

```r
mean( is.na(PM10$Site6) )
```

```
## [1] 0
```

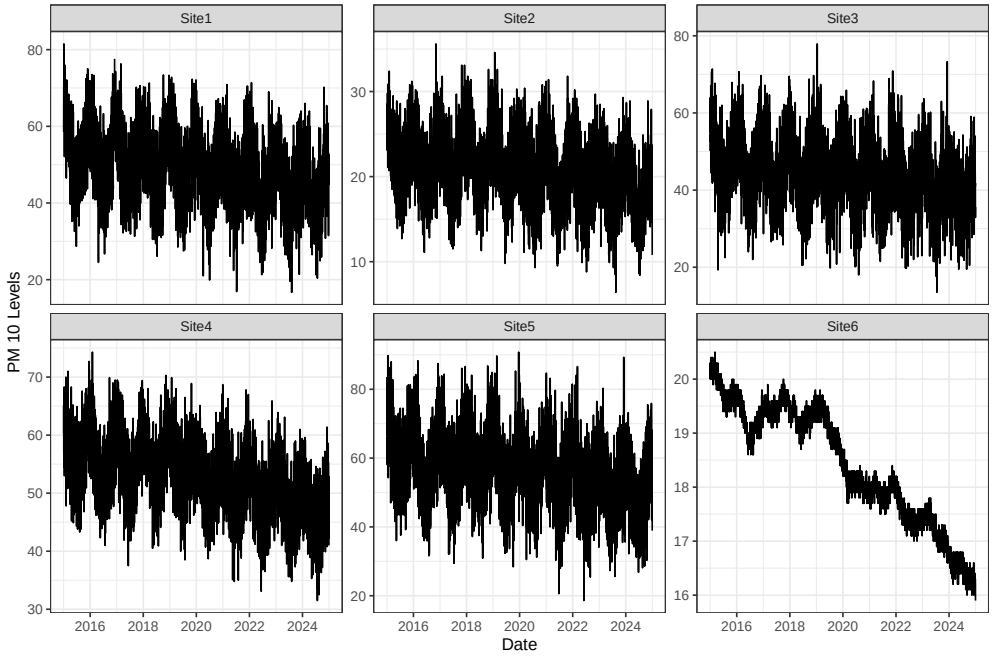
We find that 3% and 2.1% of data are missing for Site 1 and Site 4 respectively.

Our next step is to convert the variable **Date** to its correct type

```r
PM10 <- PM10 %>% mutate( Date = as_date( Date ) )
```

We now create line plots to study how PM10 levels changed over time:

```r
PM10 %>%
  pivot_longer( cols = Site1:Site6, names_to = "Site") %>%
  ggplot( aes(x=Date, y=value) ) + theme_bw() +
  facet_wrap(~Site, scales = "free_y") + geom_line() +
  labs( y="PM 10 Levels")
```



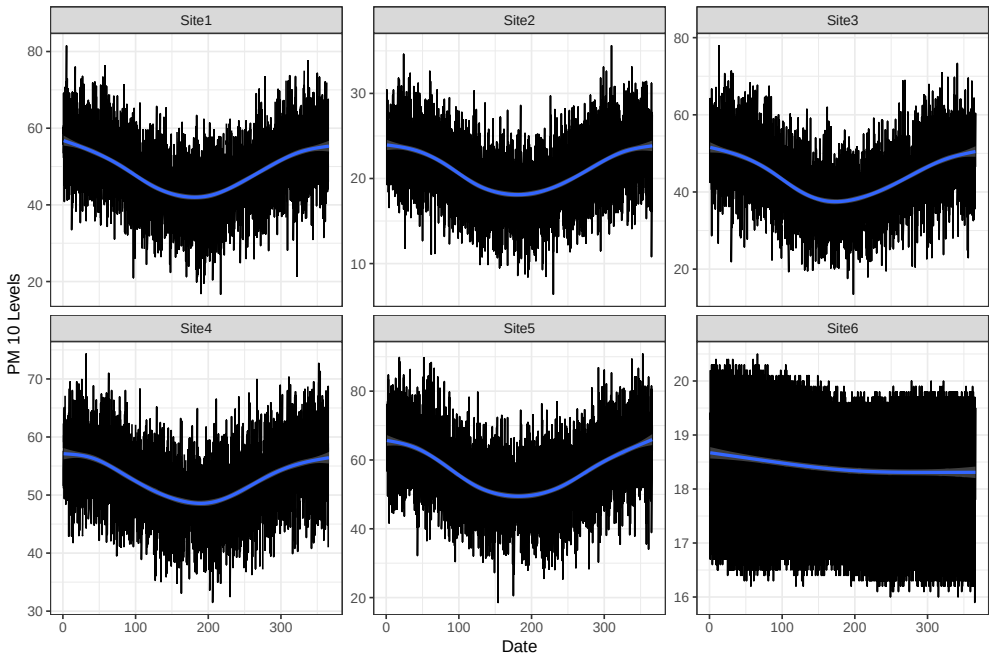
We observe:

- Possible seasonality across years
- A slight decrease in PM10 levels over time for all sites, and a stronger decrease for Site 6.
- Sites 1,3,4 and 5 observe much higher PM10 levels than Sites 2 and 6 (some more detail should be provided by the student)

To better visualize the seasonality aspect, we consider day of year

```r
PM10 %>%
  pivot_longer( cols = Site1:Site6, names_to = "Site") %>%
  mutate( Date = yday(Date) ) %>%
  ggplot( aes(x=Date, y=value) ) + theme_bw() +
```

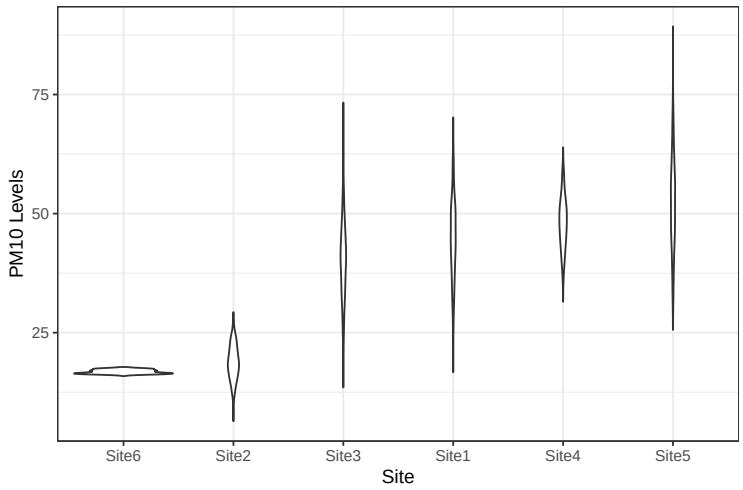
```
  facet_wrap(~Site, scales = "free_y") + geom_line() +
  geom_smooth() + labs( y="PM 10 Levels")
```



We find that PM10 levels are lower in summer than in winter. This may reflect that more PM10 emissions occur during heating season. For Site 6, there seems to be no seasonality, but the levels are quite low. So Site 6 may not be close to residential buildings. As such, it may be good to have more information on the spatial location of the sites.

Finally, we create a set of violin plots for the 2023-2024 data for the analysis (box plots are also okay, but the plots should be ordered in any case):

```
PM10 %>%
  filter( year(Date) >= 2023 ) %>%
  pivot_longer( cols = Site1:Site6, names_to = "Site") %>%
  ggplot( aes( x=reorder(Site, value, median, na.rm=TRUE), y=value ) ) +
  geom_violin() + theme_bw() +
  labs( x="Site", y="PM10 Levels" )
```



We see that Sites 1,3 and 5 are the most variable in terms of the recorded PM10 levels. Less variability is observed for Site 4, although it has the second highest median value. This may be due to a variety of reasons, such as traffic intensity, mix of industry and residential buildings, etc. So this data may be useful to have to better explain the differences we observed. Site 2 and 6 are less varied, reflecting that they have generally lower PM10 levels.

Marking guidance

To achieve 12 out of 15 marks, the solution should (i) clearly report the amount of missing data, (ii) identify the non-stationarity across the year, (iii) provide a clear summary of the distribution of values and (iv) outline two pieces of information/data that may be useful. The results for (ii) and (iii) should be supported by plots which clearly visualize the key aspects, without the number of plots being excessive or the plots being cramped. Arguments presented for (iv) should be supported by the data and not just be generic. ChatGPT lists a range of sensible answers for (iv), but they should be linked to the data. The solution above demonstrates the extent of plots that would be sufficient to achieve such a mark of 8 out of 10.

A solution may achieve 13 or more marks if it provides an in-depth discussion which sets the analysis in context, and/or which uses data visualization in an innovative way.

The following list outlines some scenarios and the mark they should achieve:

- **11 marks** - The report performs the analysis as above, but does not report the amount of missingness OR does not link suggestions for (iv) to the analysis.
- **9-10 marks** - The report addresses all aspects (i), (ii), (iii) and (iv), but the plots are not interpreted in enough detail.
- **9-10 marks** - The report performs the analysis as above, but does not report the amount of missingness AND does not link suggestions for (iv) to the analysis.
- **9-10 marks** - The report presents nice solutions for (i), (iii) and (iv), but it does not comment on the non-stationarity.
- **8 marks** - The report only addresses three of the four points, and the plots are not interpreted in enough detail.
- **6-7 marks** - Some attempt at answering the questions and the report demonstrates some understanding of how data visualization may be used to answer the question. The overall analysis lacks however depth and (i) and (iv) are barely answered.
- **at most 7 marks** - An answer that lacks any interpretation of plots and results.
- **0-5 marks** - An answer which demonstrates very little understanding of data cleaning, wrangling and visualization.

For students who achieved at least 7 marks using the guidance above, minor issues in presentation or an excessive number of plots should be penalized by deducting 1 mark. In extreme cases, such as printing the data files or providing way too many plots, students who achieved at least 8 marks will be deducted up to 2 marks.

Question 2a)

Which features in the energy consumption does the government of Utopia need to take into account when designing their new energy strategy?

Rationale

This question assesses the student's ability to effectively manipulate a dataframe and consider data at a variety of temporal scales. Specifically, a complete answer should focus on patterns throughout the day, throughout the year and other time. This also requires to aggregate observations using the dplyr R package covered in the lectures and tutorials. While there exists some connections to examples considered in lectures/problem classes/tutorials, given the variety of aspects that need to be considered, many students are expected to miss at least one key aspect.

Example of a first class solution

We start by loading the data

```
Consumption <- read.csv( "Utopia Energy Consumption.csv" )
glimpse( Consumption )
```

```
## Rows: 113,931
## Columns: 2
## $ Date   <chr> "2012-01-01 00:30:00", "2012-12-31 01:30:00", "2012-12-31 02:30~
## $ Demand <int> 12892, 14254, 13761, 13445, 13306, 13219, 13433, 13740, 14354, ~
```

We need to convert the date to the correct type

```
Consumption <- Consumption %>% mutate( Date = ymd_hms(Date) )
```

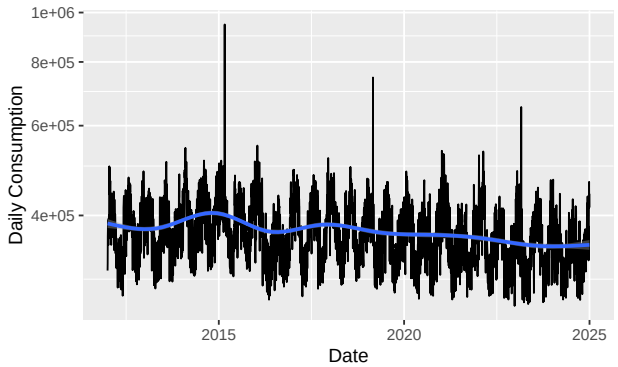
A first aspect we need to study is the variation in energy demand across the years. To do this, we consider the total consumption across a day and this is calculated using

```
Consumption_Daily <- Consumption %>%
  group_by( Date=date(Date) ) %>%
  summarise( Total = sum(Demand) )
```

We then visualize how the daily energy demand has evolved over the observation period

```
ggplot( Consumption_Daily, aes(x=Date, y=Total ) ) +
  geom_line() + geom_smooth() + coord_trans( y="log" ) +
  labs( y="Daily Consumption")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



The plot suggests that the average energy demand has dropped slightly over the past 15 years. To explore this a bit further, we calculate the average daily consumption per year

```
Consumption_Daily %>%
  group_by( Year = year(Date) ) %>%
  summarise( Total = mean(Total) )
```

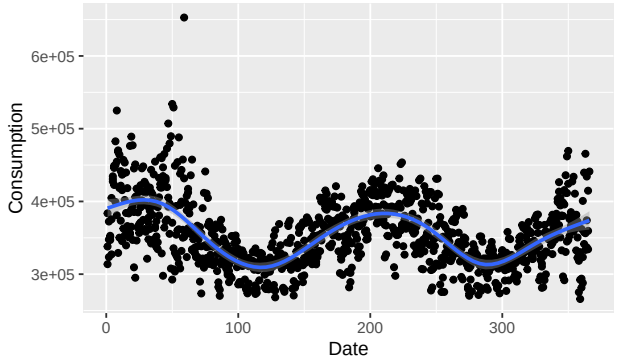
```
## # A tibble: 13 x 2
##     Year   Total
##    <dbl>   <dbl>
##  1  2012 380145.
##  2  2013 377607.
##  3  2014 399401.
##  4  2015 397876.
##  5  2016 366015.
##  6  2017 384075.
##  7  2018 379483.
##  8  2019 369354.
##  9  2020 364674.
## 10  2021 364016.
## 11  2022 356854.
## 12  2023 355794.
## 13  2024 347610.
```

We find that the daily demand has dropped from around 380000 MW in 2012 to 350000 MW in 2024.

The plot also shows some seasonal pattern, which we explore further and we only focus on the most recent years, e.g., 2022-2024:

```
Consumption_Daily %>%
  mutate( Year = year(Date) ) %>%
  filter( Year > 2021 ) %>%
  ggplot( aes(x=yday(Date),y=Total) ) +
  geom_point() + geom_smooth() +
  labs( x="Date", y="Consumption")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

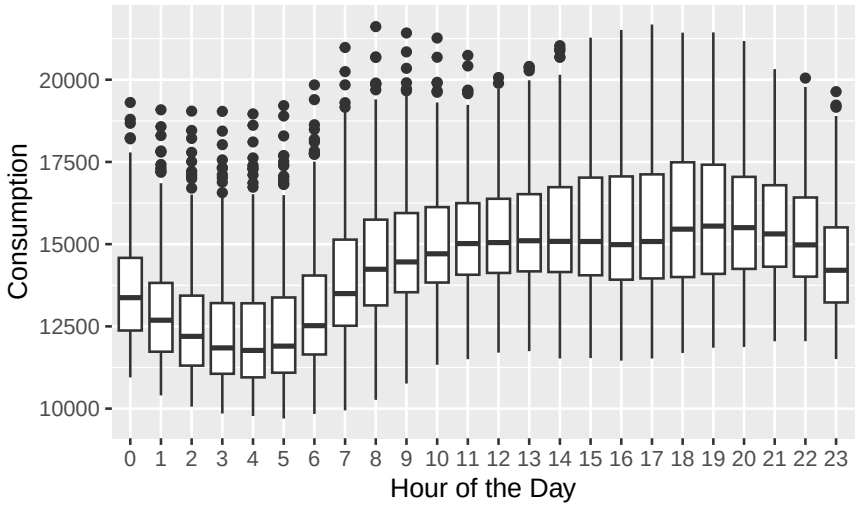


We find that energy demands peaks during summer and winter, while being lower in spring and autumn. This may reflect the effect of air condition and heating demand during periods of cold/hot weather.

Finally, we explore variation throughout the day and we just focus on last year:

```
Consumption %>%
  filter( year(Date) == 2024 ) %>%
  ggplot( aes(x=factor(hour(Date)),y=Demand) ) +
  geom_boxplot() +
```

```
labs(x="Hour of the Day", y="Consumption")
```



We find that demand is lowest in the very early morning, and it's highest in the evening around 18.00-21.00. We also see that there is a lot variation for the individual hours, which reflects the seasonal differences in demand explored before.

Marking guidance

For an answer to be considered first class, it has to explore all three aspects mentioned above: (i) change over the years, (ii) variation throughout the year and (iii) variation throughout the day. A solution may achieve 9 or 10 marks if it provides an in-depth discussion which sets the analysis in context, or presents an elegant solution in terms of data visualization, e.g.. instead of just plotting the hourly average, one may consider the hourly demand relative to daily demand.

The following list outlines some scenarios and the mark they should achieve:

- **7 marks** - The report performs the analysis as presented above, but it ignores the trend. For instance, seasonality is explored but the data for all years is considered, while it would be better to just focus on recent years.
- **6 marks** - The analysis considers two of the three aspects above and has a good structure/presentation, including clear conclusions.
- **6 marks** - The report analyses all three aspects above, but there are weaknesses in terms of interpretation and ignores the trends in the data.
- **5 marks** - The report addresses two of the three points in a reasonable manner, with minor weaknesses in the clarity of the presentation/interpretation, or ignoring the trend.
- **4 marks** - One of the three aspects has been studied in a nice level of detail.
- **4 marks** - The report addresses two of the three points, but there are some major weaknesses in terms of presentation and interpretation.
- **at most 4 marks** - An answer that lacks any interpretation of plots and results.
- **0-3 marks** - An answer which demonstrates very little understanding of data cleaning, wrangling and visualization.

Question 2b)

Energy in Utopia is produced using four sources: several solar farms, wind farms, coal-fired power stations and a nuclear power station. Explore how much renewable energy sources, solar and wind, have been contributing to the energy production.

The current capacity for energy production using coal is 900,000 MW. An advisor to the government of Utopia suggests to reduce this capacity by 40% and to close the nuclear power station within the next two years. The latter is motivated by the country's difficulty to find a secure ultimate disposal place for the nuclear waste. What would be the impact of this plan on the country's capacity to meet its energy demand?

Do you have any suggestions on how this plan may be improved? In your discussion, please take into account that Utopia is unable to import energy from abroad and to immediately build new solar and wind farms.

Rationale

This question assesses the student's ability to work with more thna one data frame for one analysis. The first part can be answered by using a single data frame, but the remaining aspects require the combination of information from two data sets, for instance, using the dplyr R package. There is also the task for presenting the results in a clear, concise and structured manner.

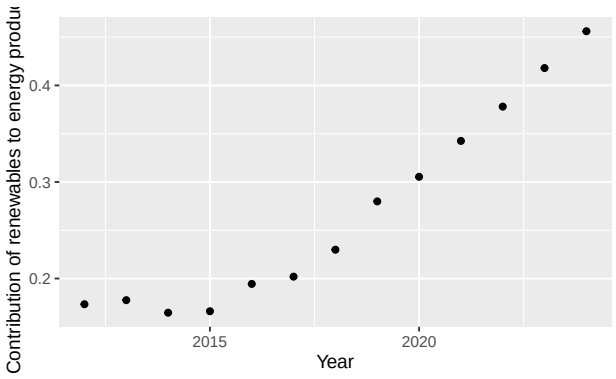
Example of a first class solution

We start by considering how much renewable energy sources have contributed to Utopia's energy production and load the relevant data and convert to the correct type

```
Production <- read.csv("Utopia Energy Production.csv")
Production <- Production %>% mutate( Day = as_date(Day) )
```

There are two aspects that need to be considered. First we consider how much solar and wind contributed to energy production over the year and we study whether there are seasonal differences - energy produced from solar farms is dependent on the amount of daylight. We start by considering production per year and visualize it using, for instance, a scatter plot

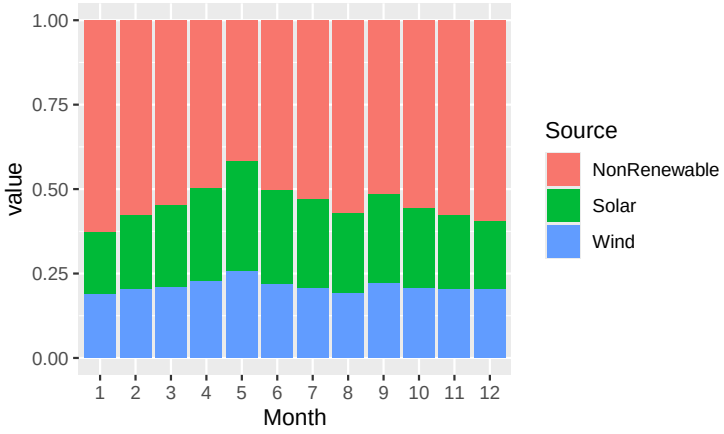
```
Production %>%
  group_by( Year = year(Day) ) %>%
  summarise( Renewable=sum(Solar+Wind),
             Total = sum(Solar+Wind+Coal+Nuclear) ) %>%
  mutate( Proportion = Renewable / Total) %>%
  ggplot( aes(x=Year, y=Proportion) ) + geom_point() +
  labs( x="Year", y="Contribution of renewables to energy production")
```



We see that renewable energy sources contributed less than 20% until 2015, but their proportion has increased since to up to around 45% in 2024. As a next step, we now consider differences throughout the year, and we

only consider the year 2024 as it is the most relevant for developing the government's new strategy. We also consider solar and wind separately as it may provide more insight:

```
Production %>%
  filter( year(Day) == 2024 ) %>%
  mutate( Month = month(Day) ) %>%
  group_by( Month ) %>%
  summarise( Solar=sum(Solar), Wind=sum(Wind),
             NonRenewable = sum(Coal+Nuclear) ) %>%
  pivot_longer( cols = Solar:NonRenewable, names_to = "Source" ) %>%
  ggplot( aes(x=factor(Month),y=value, fill = Source ) ) +
  geom_bar( stat="identity", position = "fill" ) +
  labs( x="Month", ylab="Contribution to Energy Production" )
```



We see that there are some differences between the months, with solar farms contributing more in the summer than the winter months. Further, in the summer, renewable energy sources account for 50% of Utopia's energy production.

We now move on analysing the impact of the proposal. Let's take energy production 2024 as our data set. Under this proposal, energy production from coal cannot exceed 40% of the current capacity of 500,000 MW, that is 300000, and output from the nuclear power station is equal to zero. Let's update the amount of produced energy we would have seen:

```
Production_2024_alt <- Production %>%
  filter( year(Day) == 2024 ) %>%
  mutate( Nuclear = 0,
          Coal = case_when( Coal>300000 ~ 300000, .default = Coal) ) %>%
  mutate( Total = Coal + Solar + Wind + Nuclear )
```

We now need to relate this to the daily demand, adn thus we need to combine this data with that for energy demand. Using the same steps as for Question 2a), we have

```
Consumption <- read.csv( "Utopia Energy Consumption.csv" )
Consumption_Daily <- Consumption %>%
  mutate( Date = ymd_hms(Date) ) %>%
  group_by( Date=date(Date) ) %>%
  summarise( Demand = sum(Demand) )
```

We now extract the data for 2024 and combine it with the data frame for Production

```
Energy2024_alt <- Consumption_Daily %>%
  filter( year(Date)==2024 ) %>%
  full_join( Production_2024_alt, by = c("Date"="Day") )
```

Now we can explore the proportion of days on which demand exceeded the amount of energy produced

```
Energy2024_alt %>%
  summarise( Number = sum( Demand > Total),
             Proportion = mean( Demand > Total) )
```

```
## # A tibble: 1 x 2
##   Number Proportion
##    <int>      <dbl>
## 1     29     0.0795
```

So we find that there would be 29 days on which demand exceeded the amount of energy produced. Consequently, there seems to be a massive risk with the proposal which may lead to blackouts on multiple across the year. We can study the impact in more detail by looking at the months most affected:

```
Energy2024_alt %>%
  group_by( Month=month(Date) ) %>%
  summarise( Number = sum( Demand > Total),
             Proportion = mean( Demand > Total) ) %>%
  slice_max( Number, n=4 )
```

```
## # A tibble: 4 x 3
##   Month Number Proportion
##   <dbl>  <int>      <dbl>
## 1     8      7      0.226
## 2     1      5      0.161
## 3    10      5      0.161
## 4    12      4      0.129
```

So there are three months (which are spread randomly across the year) for which demand would exceed production on five or more days. Consequently, switching off the nuclear power station and some of the coal-based power plants immediately seems not advisable.

Should the increase in energy produced from renewable energy sources continue, the risk of blackouts would be much lower, and thus the government of Utopia may want to carefully begin this process. For instance, if production from coal was capped at 70% of current capacity as proposed, and the nuclear power station continued for a little longer, we find that demand would never have exceeded production in 2024:

```
Production_2024_alt <- Production %>%
  filter( year(Day) == 2024 ) %>%
  mutate( Coal = case_when( Coal>350000 ~ 350000, .default = Coal) ) %>%
  mutate( Total = Coal + Solar + Wind + Nuclear )

Consumption_Daily %>%
  filter( year(Date)==2024 ) %>%
  full_join( Production_2024_alt, by = c("Date"="Day") ) %>%
  summarise( Number = sum( Demand > Total),
             Proportion = mean( Demand > Total) )
```

```
## # A tibble: 1 x 2
##   Number Proportion
##    <int>      <dbl>
## 1      0          0
```

Consequently, the government could start reducing their reliance on coal now, and slowly transition away from coal and nuclear as solar and wind increase their production.

Marking guidance

For an answer to be awarded a first class mark it should (i) consider the two aspects related to energy production, (ii) identify that the proposal has risks, (iii) propose one sensible alternative and (iv) states at least one core assumption underlying the analysis. An analysis as outlined above may be awarded a mark of 12 out of 15. There is a lot of potential for innovative approaches when it comes to discussing the impacts of the proposed or some very nice discussions on the assumption. Such advanced aspects may fetch a mark above 80%.

The following list outlines some scenarios and the mark they should achieve:

- **11 marks** - The report performs the analysis as presented above, but there are some weaknesses in the argument.
- **9-10 marks** - The analysis considers the key aspects, but critical assumptions have not been identified.
- **9-10 marks** - The answer only considers one of the two aspects on energy production, but addresses all other points in a suitable way.
- **8 marks** - The report makes a decent attempt at addressing all the questions, but there are weaknesses in the quality of the analysis and the level of explanations.
- **6-7 marks** - Some attempt has been made, but the arguments are flawed and presented poorly.
- **at most 5 marks** - An answer that lacks any interpretation of plots and results.
- **0-5 marks** - An answer which demonstrates very little understanding of data cleaning, wrangling and visualization.