# FIT5145 S1 2019 Assignment 3

**Semester 1, 2019**

Student Name: Anik Dey Sarker

## Part A: Investigating the data in the Shell

Download the file ELEC.zip. Use a Unix shell to manipulate the file and answer the following questions.

1) Decompress the file ELEC.zip. How big is the file ELEC.txt that is obtained after unzipping?

   **Linux command:** **unzip ELEC.zip**
   **ls -l ELEC.txt --block-size=MB**

   **Answer: to decompress the file ELEC.zip I've used `unzip ELEC.zip` command. To view how big is the file ELEC.txt I've used `ls -l ELEC.txt --block-size=MB` command and the size of ELEC.txt file is 954 MB.**

**2)** Based on visual inspection of parts of the file what is the most common delimiter used in the file? What units are used to quantify "electric fuel consumption"**?**

   **Linux command:** **head -1 ELEC.txt | less**

   **Answer: The most common delimiter used is ' , '. I used "/," to find delimiter. The unit used to quantity "electric fuel consumption" is "MMBtu".**

**3)** Almost every line in the file that begins with the field "series_id" provides data on "electric fuel consumption" for a specific power plant in the USA. Apart from the field "series_id", what are the other fields provided in a line containing the field "series id"?

   **Answer:**
**'name,units,f,description,copyright,source,iso3166,lat,lon,geography,start,end ,last updated,latlon,data'**

**4)** How many lines are there in the file?

   **Linux command:** **wc -l ELEC.txt**

   **Answer: There are total 609571 lines in the file.**

5) For each line containing the field "series_id" and a given powerplant, what does the field "f" represent? What is the date range that the monthly electric fuel consumption data spans for the first power plant in the file ("Arlington Wind Power Project (56855)" for "all fuels" and "all primemovers") and for which there is actually electric fuel consumption data?

**Linux command:** **head -1 ELEC.txt | less**

**Answer:** **“f” represents time span of fuel consumption like ‘M’ for Monthly fuel consumption, ‘Q’ for Quarterly fuel consumption or ‘A’ Annual fuel consumption . The date range for the first power plant is from December 2008 to December 2018.**

6) How many lines in the file contain the field “series_id”?

**Linux command:** **grep 'series_id' ELEC.txt | wc -l**

**Answer:** **There are 592745 lines which contain the field “series_id”.**

7) How many unique power plants are named in the file in the lines containing the field “series_id”? Note that some power plants occur on multiple lines based on different information provided for a given power plant on a given line containing the field “series_id”.

**Linux command:** **grep 'series_id’ ELEC.txt | awk '{print substr($0,2,length()-2);}' | cut -d ',' -f 2 | cut -d ':' -f 3 | uniq | wc**

**Answer:** **There are 187545 unique power plants named in the lines containing the field “series_id”.**

**8)** On which month and year was “electric fuel consumption” the highest for the “12 Applegate Solar LLC (59371)” power plant when considering “solar” fuel and “all primemovers”? What was the amount of electric fuel consumption in at this time? (Hint: “electric fuel consumption” data is captured in the “data” field)

**Linux command:** **grep 'Electric fuel consumption MMBtu : 12 Applegate Solar LLC (59371) : solar : all primemovers : monthly' ELEC.txt | awk '{print substr($0,2,length()-2);}' | cut -d ',' -f 18- | cut -d ':' -f 2 | awk '{print substr($0,2,length()-2);}' | tr '[]' ' ' | tr , '\n' | paste -d ' ' - - | sort -k2 -nr | head -1**

**Answer:** **In April 2013, electric fuel consumption was the highest for the ‘12 Applegate Solar LLC (59371)’ considering ‘solar’ fuel and ‘all primemovers’, and the amount was 2687 MMBtu.**

9) How many times has the “126 Grove Solar LLC (60858)” power plant been listed in the file? Is this number equal to the number of lines containing “126 Grove Solar LLC (60858)” in the file?

**Linux command:** **grep “126 Grove Solar LLC (60858)” ELEC.txt | wc**
**grep “126 Grove Solar LLC (60858)” ELEC.txt | wc -l**

**Answer: “126 Grove Solar LLC (60858)” power plant has been listed 28 times in the file, and it is equal to the number of lines containing “126 Grove Solar LLC (60858)” in the file.**

10) Do you think we would be able to compute correlations (e.g. Pearson’s correlation) in electric fuel consumption between power plants using the data provided here? What problems might we face in doing so? If instead we wanted to make predictions about tomorrow’s electric fuel consumption at a given power plant, what problems would we face?

**Answer: I do not think it is possible to compute correlations in electric fuel consumption between power plants because to compute correlations, the two variables must be continuous.**

**In the case of predicting tomorrow’s electric fuel consumption at a given power plant, there is not enough predictor variables in the data. We can only plot fuel consumption against time which may or may not provide any valuable insight.**

## Part B: Graphing the data in R

1) We want to visualize, analyse and make future predictions about the “electric fuel consumption” for the “12 Applegate Solar LLC (59371)” power plant when considering “solar” fuel and “all primemovers”. First we need to extract the data corresponding with the “data” field using Bash so that we can save it as the file “fuel_data.txt” and load it into R. Provide the Bash command you used to do this and provide a table with two columns (date, electric fuel consumption) containing the annual values of electric field consumption for the month of December. (Note 1: the “fuel_data.txt” file should contain the data for all months in order to answer the other remaining questions for this part of the assignment – we only ask for the December values here since it is easier to mark; Note 2: remember to submit the “fuel_data.txt” file with your assignment).

**Linux Command:** **grep 'Electric fuel consumption MMBtu : 12 Applegate Solar LLC (59371) : solar : all primemovers : monthly' ELEC.txt | awk '{print substr($0,2,length()-2);}' | cut -d ',' -f 18- | cut -d ':' -f 2 | sed -e 's/\[//g' -e 's/\]//g' -e 's/\"//g' -e 's/,/\n/g' | paste -d ' ' - - | sed 's/./&\//4;s/./&\/1/7' | sed '1 s/^/date \"fuel consumption\"\n/' > fuel_data.txt**

**submitted file:** **fuel_data.txt**

**Answer:**

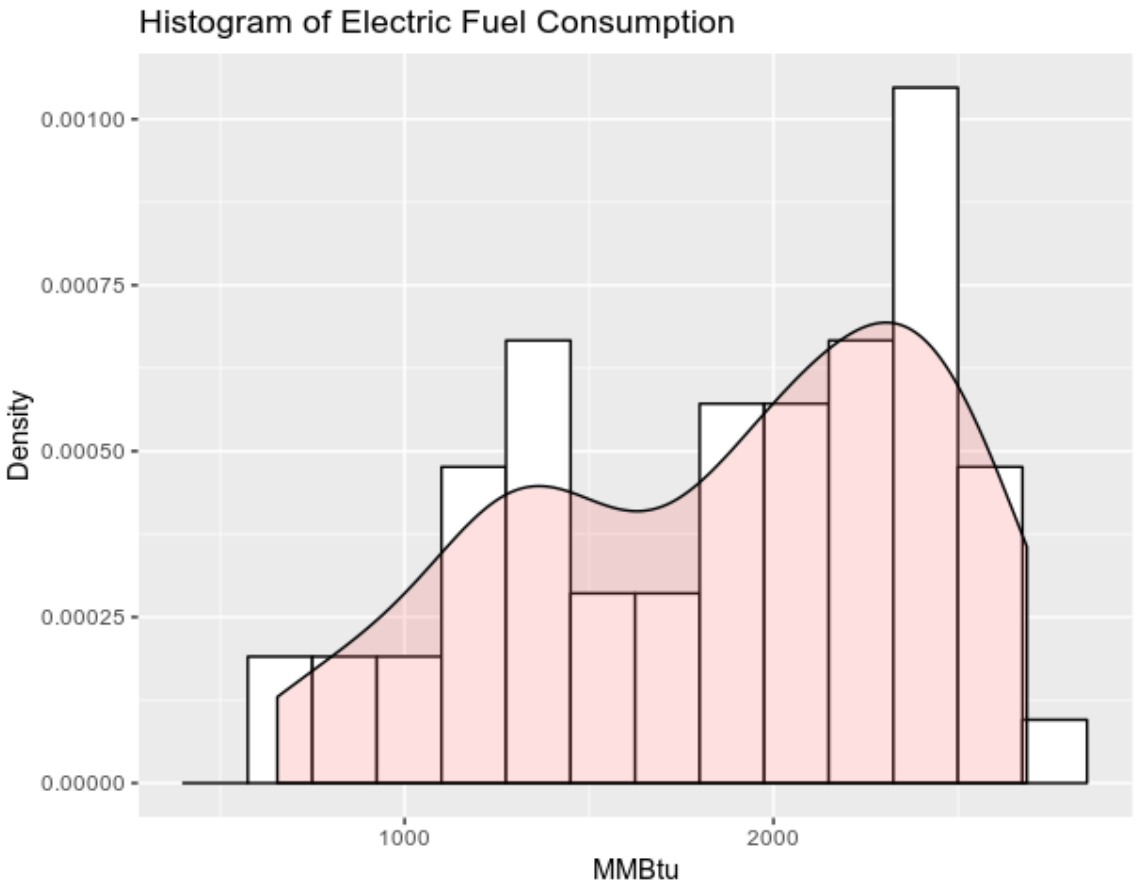
| Date | Electric Fuel Consumption (MMBtu) |
|---|---|
| 2017-12-01 | 728 |
| 2016-12-01 | 1156 |
| 2015-12-01 | 764 |
| 2014-12-01 | 922 |
| 2013-12-01 | 958 |

2) Load in the file you created “fuel_data.txt” and plot a histogram of the electric fuel consumption with labels on the axes and a title. Does the data follow a Gaussian distribution?

**R scripts:**

```
> data <- read.table('~/fuel_data.txt', header = TRUE, sep = " ")
> ggplot(data=data, aes(x = fuel.consumption)) + geom_histogram(breaks = seq(400, 3000, by = 175), col = 'black', fill = 'white', aes(y=..density..)) + geom_density(alpha=.2, fill="#FF6666") + labs(title = 'Histogram of Electric Fuel Consumption') + labs(x = 'MMBtu', y = 'Density')
```

**Plot:**



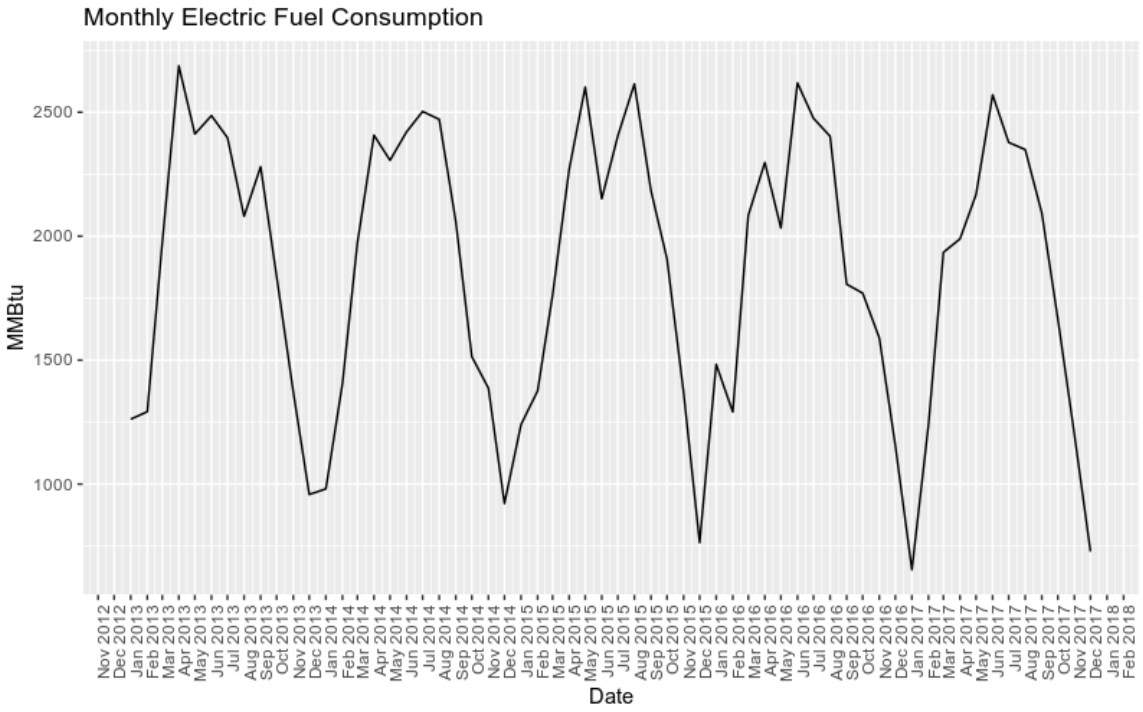

**Answer:** **As it is evident from the plot, the distribution is not bell shaped, and it is not a Gaussian distribution.**

3) Now plot the monthly electric fuel consumption data as a function of time with time increasing in the rightward direction of the plot and with labels on the axes and a title. Based on looking at this time series give a reason why you gave your answer in question 2 above as to whether or not the electric fuel consumption data followed a Gaussian distribution. What issues would you face if you tried to fit a linear regression to this data to predict fuel consumption during times occurring just after the data provided?

**R scripts:**

```
> ggplot(data, aes(date, fuel.consumption)) + geom_line() +
scale_x_date(date_labels = "%b %Y", date_breaks = "1 month") +
theme(axis.text.x = element_text(angle = 90, hjust = 1)) + labs(x = "Date", y =
"MMBtu", title = "Monthly Electric Fuel Consumption")
```

**Plot:**




**Answer: The graph is not a straight line, so linear regression can not be used to best fit this model. Rather, a polynomial regression might be used to predict future consumptions.**

## Part C: Investigating Chronic Disease Indicators Data

Download the file U.S._Chronic_Disease_Indicators__CDI_.zip. Use a Unix shell to manipulate the file and answer the following questions.

1) Decompress the file U.S._Chronic_Disease_Indicators__CDI_.zip. Unzipping this file outputs U.S._Chronic_Disease_Indicators__CDI_.csv. View the start of this csv file. What information does the first line in the file provide about the remaining lines in the file?

   **Linux command: unzip U.S._Chronic_Disease_Indicators__CDI_.zip**
   **head -1 U.S._Chronic_Disease_Indicators__CDI_.csv**

   **Answers: The first line of the file is a header line that provides information about the respective columns.**

2) The file provides information about Chronic Diseases in the United States by referring to disease “topic”s and “questions” (i.e. indicator measures) that relate to these topics. Use an awk script to extract only the lines providing “Crude Prevalences” for disease topic questions for the state of California and save them to a file called ‘california.txt’.

**Linux command:**

**awk -F, '{if($4 == "California" && $10 == "Crude Prevalence") {print}}' U.S._Chronic_Disease_Indicators__CDI_.csv > california.txt**

**submitted file: california.txt**

3) Considering the new file “california.txt”:
   A. How many lines are associated with the disease “topic”s “Alcohol” and “Cancer”?

**Linux command: awk -F, '{if ($6 == "Alcohol" || $6 == "Cancer") {print}}' california.txt | wc**

**Answers: There are 263 lines are associated with the disease “topic”s “Alcohol” and “Cancer”**

   B. What is the highest Crude Prevalence value for the “Cancer” “topic” in this file? What “year” and “question” is associated with this highest crude prevalence value for “Cancer”? What type of cancer does this “question”/indicator measure most likely relate to? Why would Crude Prevalence be high for this “question”/indicator measure?

**Linux command: awk -F, '{if ($6 == "Cancer") {print}}' california.txt | sort -t, -gk11**

**Answers: The highest Crude Prevalence value for the “Cancer” topic in this file is 91.4. The associated year is 2014 and the question is about “Mammography use among women aged 50-74 years. As the question is about Mammography, it is most likely related to breast cancer. Crude prevalence is high for this question because breast cancer is the most common cancer diagnosed in women.**