**Electric Vehicle Charging Dataset**

Data from 85 EV drivers with repeat usage at 105 stations across 25 sites

**About Dataset**

This dataset contains information from 3,395 high resolution electric vehicle charging sessions. The data contains sessions from 85 EV drivers with repeat usage at 105 stations across 25 sites at a workplace charging program. The workplace locations include facilities such as research and innovation centers, manufacturing, testing facilities and office headquarters for a firm participating in the U.S. Department of Energy (DOE) workplace charging challenge. The data is in a human and machine readable \*.CSV format.

Here is a list of the self-explained variables:

sessionId

kwhTotal

created

ended

chargeTimeHrs

weekday

platform

userId

stationId

managerVehicle

**Problem 1 – Loading the Data**

Read the dataset into R. Use the R functions to answer the following questions.

1.1 How many rows of data are in this dataset?

|  |  |
| --- | --- |
| Code: data<-read.csv("C:/Users/humza/OneDrive/Desktop/station\_data\_dataverse.csv")  nrow(data) | Output: |

1.2 What are the datatypes of each variable in the dataset? How many missing / NA values in each variable.

|  |  |  |
| --- | --- | --- |
| Variable Name | DataTypes | No of Missing/NA Values |
| sessionId |  |  |
| kwhTotal |  |  |
| created |  |  |
| ended |  |  |
| chargeTimeHrs |  |  |
| weekday |  |  |
| platform |  |  |
| userId |  |  |
| stationId |  |  |
| managerVehicle |  |  |

1.3 For each variable, design an appropriate strategy to fill the missing/NA values. Provide rationale and code for filling the missing values. Clearly, mention the tradeoffs and assumptions that you will consider for this task.

|  |
| --- |
| summary(data)  #it tells us alot about the data even the minute details in which median is one of them  #we are using median as mean is simply effected by outliers for replacing na values.  data$chargeTimeHrs[is.na(data$chargeTimeHrs)]=2.8057  data$kwhTotal[is.na(data$kwhTotal)]=6.210 |

**Problem 2 – Understanding Dates in R**

In many datasets, like this one, you have a date field. Unfortunately, R does not automatically recognize entries that look like dates. We need to use a function in R to extract the date and time. Look at “created” and “ended” entries of Dates

2.1 In what format are the entries in the variable Date?

|  |  |
| --- | --- |
| Code: str(data$created) | Output: |

2.2 convert these characters into a Date object in R.

|  |  |
| --- | --- |
| Code: as.Date(data$created) | Output: |

2.3 What is the month and year of the median date in our dataset? Enter your answer as "Month Year", without the quotes. (Ex: if the answer was 2008-03-28, you would give the answer "March 2008", without the quotes.)

|  |  |
| --- | --- |
| Code: med<-median(data$kwhTotal)  d<-(data[data$kwhTotal== med,])  d  dat<-lubridate::as\_date(d[1,5])  format(dat,"%m")  m<-lubridate::month(dat,label = TRUE)  y<-format(dat,"%Y")  final<-paste(m,y) | Output: |

2.4 What is the weekday of the median date in our dataset? Enter your answer as "Weekday", without the quotes.

|  |  |
| --- | --- |
| Code: week<-(d[1,6])  weekday<-gsub(" ","",paste(week,"day"))  weekday | Output: |

2.5 On which weekday did the most recharging of vehicle occurs?

|  |  |
| --- | --- |
| Code: d<-(data[data$kwhTotal== (max(data$kwhTotal,na.rm = TRUE)),])  week<-(d[1,6])  weekday<-gsub(" ","",paste(week,"day"))  weekday | Output: |

**Problem 3 – Visualizing the Charging Trends**

Let's investigate this further. User plotting answer the following questions

3.1 In general, are no of charging at any station increases or decreases on daily basis?

|  |  |
| --- | --- |
| Code: data$stationId[1]  data[data$stationId==data$sessionId[4],]  library(ggplot2)  ggplot(data)+  geom\_point(aes(weekday,chargeTimeHrs,color=stationId)) | Output: |

3.2 Which are the top 10 popular stations?

|  |  |
| --- | --- |
| Code: ggplot(data)+  geom\_line(aes(stationId,chargeTimeHrs))  count\_data<-data %>%  count(stationId,name="count")  ggplot(data)+  geom\_bar(aes(stationId))  ggplot(data) +  geom\_bar(aes(x = stationId, fill = as.factor(userId)),  position = "dodge", order(stat = "count"),10) +  facet\_wrap(~stationId) | Output: |

3.3 How many observations are in these top 10 stations?

|  |  |
| --- | --- |
| Code: | Output: |

3.4 Which fueling stations are the most economical ones? (Hint: ratio between kwhTotal and chargeTime).

|  |  |
| --- | --- |
| Code: | Output: |

3.5 Is there any difference of charging behavior on manager vehicles than non-manager vehicle ones?

|  |  |
| --- | --- |
| Code: ggplot(data)+  geom\_point(aes(managerVehicle,chargeTimeHrs)) | Output: |