VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

CSE3046-Programming for Data Science

Process: Obtain Data from Various Resources

Assignment-3

HANDLING MISSING DATA

NAME : D VASANTH KUMAR

REG NO: 19BDS0083

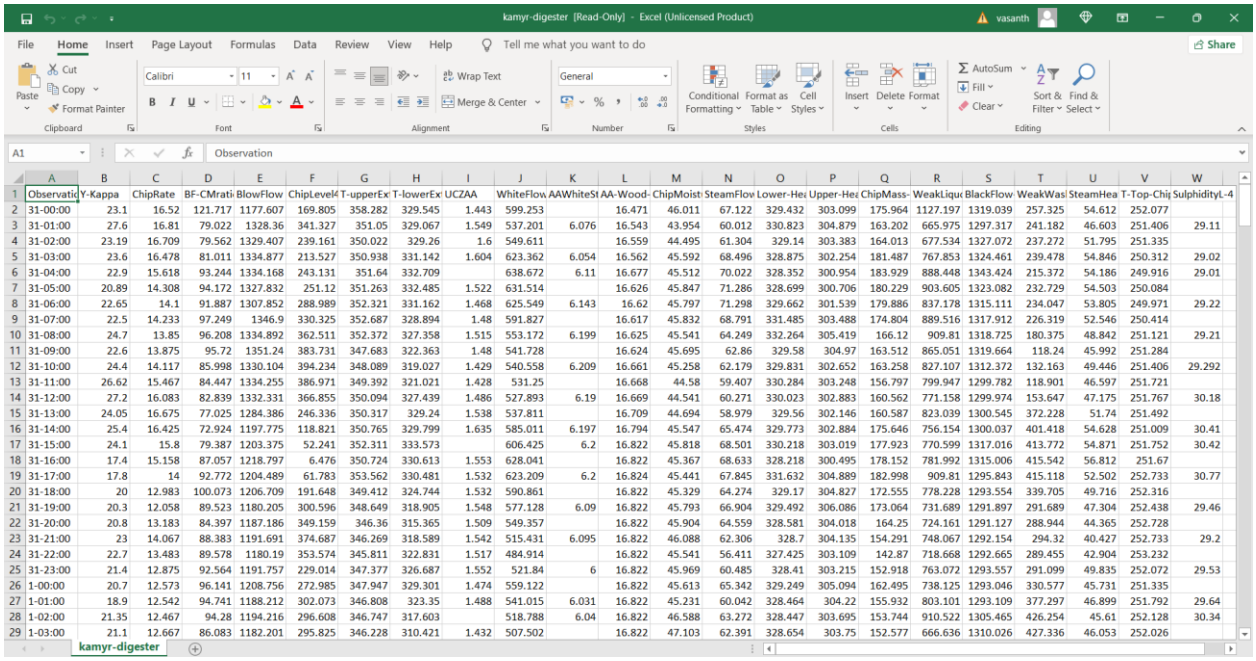
Submitted Date: 04/10/2021

Course Instructor: Dr. Anthoniraj A

Retrieve the data set from the following URL,

https://openmv.net/file/kamyr-digester.csv

THE DATA SET:



kamyr-digester [Read-Only] - Excel (Unlicensed Product)

Observation	Y-Kappa	ChipRate	BF-CMrati	BlowFlow	ChipLevel4	T-upperEx	T-lowerEx	UCZAA	WhiteFlow	AAWhiteSt	AA-Wood-	ChipMoist	SteamFlow	Lower-Hea	Upper-Hea	ChipMass-	WeakLiqu	BlackFlow	WeakWas	SteamHea	T-Top-Chi	SulphidityL-4
31-00:00	23.1	16.52	121.717	1177.607	169.805	358.282	329.545	1.443	599.253		16.471	46.011	67.122	329.432	303.099	175.964	1127.197	1319.039	257.325	54.612	252.077	
31-01:00	27.6	16.81	79.022	1328.36	341.327	351.05	329.067	1.549	537.201	6.076	16.543	43.954	60.012	330.823	304.879	163.202	665.975	1297.317	241.182	46.603	251.406	29.11
31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.26	1.6	549.611		16.559	44.495	61.304	329.14	303.383	164.013	677.534	1327.072	237.272	51.795	251.335	
31-03:00	23.6	16.478	81.011	1334.877	213.527	350.938	331.142	1.604	623.362	6.054	16.562	45.592	68.496	328.875	302.254	181.487	767.853	1324.461	239.478	54.846	250.312	29.02
31-04:00	22.9	15.618	93.244	1334.168	243.131	351.64	332.709		638.672	6.11	16.677	45.512	70.022	328.352	300.954	183.929	888.448	1343.424	215.372	54.186	249.916	29.01
31-05:00	20.89	14.308	94.172	1327.832	251.12	351.263	332.485	1.522	631.514		16.626	45.847	71.286	328.699	300.706	180.229	903.605	1323.082	232.729	54.503	250.084	
31-06:00	22.65	14.1	91.887	1307.852	288.989	352.321	331.162	1.468	625.549	6.143	16.62	45.797	71.298	329.662	301.539	179.886	837.178	1315.111	234.047	53.805	249.971	29.22
31-07:00	22.5	14.233	97.249	1346.9	330.325	352.687	328.894	1.48	591.827		16.617	45.832	68.791	331.485	303.488	174.804	889.516	1317.912	226.319	52.546	250.414	
31-08:00	24.7	13.85	96.208	1334.892	362.511	352.372	327.358	1.515	553.172	6.199	16.625	45.541	64.249	332.264	305.419	166.12	909.81	1318.725	180.375	48.842	251.121	29.21
31-09:00	22.6	13.875	95.72	1351.24	383.731	347.683	322.363	1.48	541.728		16.624	45.695	62.86	329.58	304.97	163.512	865.051	1319.664	118.24	45.992	251.284	
31-10:00	24.4	14.117	85.998	1330.104	394.234	348.089	319.027	1.429	540.558	6.209	16.661	45.258	62.179	329.831	302.652	163.258	827.107	1312.372	132.163	49.446	251.406	29.292
31-11:00	26.62	15.467	84.447	1334.255	386.971	349.392	321.021	1.428	531.25		16.668	44.58	59.407	330.284	303.248	156.797	799.947	1299.782	118.901	46.597	251.721	
31-12:00	27.2	16.083	82.839	1332.331	366.855	350.094	327.439	1.486	527.893	6.19	16.669	44.541	60.271	330.023	302.883	160.562	771.158	1299.974	153.647	47.175	251.767	30.18
31-13:00	24.05	16.675	77.025	1284.386	246.336	350.317	329.24	1.538	537.811		16.709	44.694	58.979	329.56	302.146	160.587	823.039	1300.545	372.228	51.74	251.492	
31-14:00	25.4	16.425	72.924	1197.775	118.821	350.765	329.799	1.635	585.011	6.197	16.794	45.547	65.474	329.773	302.884	175.646	756.154	1300.037	401.418	54.628	251.009	30.41
31-15:00	24.1	15.8	79.387	1203.375	52.241	352.311	333.573		606.425	6.2	16.822	45.818	68.501	330.218	303.019	177.923	770.599	1317.016	413.772	54.871	251.752	30.42
31-16:00	17.4	15.158	87.057	1218.797	6.476	350.724	330.613	1.553	628.041		16.822	45.367	68.633	328.218	300.495	178.152	781.992	1315.006	415.542	56.812	251.67	
31-17:00	17.8	14	92.772	1204.489	61.783	353.562	330.481	1.532	623.209	6.2	16.824	45.441	67.845	331.632	304.889	182.998	909.81	1295.843	415.118	52.502	252.733	30.77
31-18:00	20	12.983	100.073	1206.709	191.648	349.412	324.744	1.532	590.861		16.822	45.329	64.274	329.17	304.827	172.555	778.228	1293.554	339.705	49.716	252.316	
31-19:00	20.3	12.058	89.523	1180.205	300.596	348.649	318.905	1.548	577.128	6.09	16.822	45.793	66.904	329.492	306.086	173.064	731.689	1291.897	291.689	47.304	252.438	29.46
31-20:00	20.8	13.183	84.397	1187.186	349.159	346.36	315.365	1.509	549.357		16.822	45.904	64.559	328.581	304.018	164.25	724.161	1291.127	288.944	44.365	252.728	
31-21:00	23	14.067	88.383	1191.691	374.687	346.269	318.589	1.542	515.431	6.095	16.822	46.088	62.306	328.7	304.135	154.291	748.067	1292.154	294.32	40.427	252.733	29.2
31-22:00	22.7	13.483	89.578	1180.19	353.574	345.811	322.831	1.517	484.914		16.822	45.541	56.411	327.425	303.109	142.87	718.668	1292.665	289.455	42.904	253.232	
31-23:00	21.4	12.875	92.564	1191.757	229.014	347.377	326.687	1.552	521.84	6	16.822	45.969	60.485	328.41	303.215	152.918	763.072	1293.557	291.099	49.835	252.072	29.53
1-00:00	20.7	12.573	96.141	1208.756	272.985	347.947	329.301	1.474	559.122		16.822	45.613	65.342	329.249	305.094	162.495	738.125	1293.046	330.577	45.731	251.335	
1-01:00	18.9	12.542	94.741	1188.212	302.073	346.808	323.35	1.488	541.015	6.031	16.822	45.231	60.042	328.464	304.22	155.932	803.101	1293.109	377.297	46.899	251.792	29.64
1-02:00	21.35	12.467	94.28	1194.216	296.608	346.747	317.603		518.788	6.04	16.822	46.588	63.272	328.447	303.695	153.744	910.522	1305.465	426.254	45.61	252.128	30.34
1-03:00	21.1	12.667	86.083	1182.201	295.825	346.228	310.421	1.432	507.502		16.822	47.103	62.391	328.654	303.75	152.577	666.636	1310.026	427.336	46.053	252.026	

kamyr-digester

ACTIVITY DESCRIPTION:

- Read the data set from the URL and converting it into data frame
- summarize the data set for checking number of NA's
- check for structural errors
- Replace empty cells with NA
- Remove the columns and rows having missing data more than 60%
- Apply Mean or Median for Imputing Missing Values in remaining columns
- Apply any other standard algorithm for imputing missing values (for at least one column).

SAMPLE CODE:

```
#-------------------------------------------------------------------------------------------------------------------------------
#WRITING META DATA

#USER INFROMATION : 19BDS0083 D VASANTH KUMAR

#DATA SOURCE : Kamyr digester

#OWNED BY: OpenMV.net Datasets

#Description:Pulp quality is measured by the lignin content remaining in the pulp: the Kappa number. This data set is used to understand
            # which variables in the process influence the Kappa number,and if it can be predicted accurately enough for an
            # inferential sensor application.Variables with a number at the end have been lagged by that number of hours to line up the data.

#Data shape:    301 rows and 22 columns
#TAGS FOR THE DATA SET : MULTIVARIATE,MISSING-DATA,TIME-SERIES

#-------------------------------------------------------------------------------------------------------------------------------

#LIBRARIES USED
library(rvest)
library(dplyr)
library(tidyr)
library(utils)

#DETERMINING THE SIZE OF THE DATA SET IN MY DIRECTORY

file.info("C:/Users/vasanth kumar/Downloads/kamyr-digester.csv")$size

#READING THE DATA SET

df <-read.csv("https://openmv.net/file/kamyr-digester.csv")
#STORING THE DATA SET IN NEW VARIABLE KEEPS THE ORIGINALITY OF THE DATA SET BEFORE CLEANING
df1 <-data.frame(df)

#AN INITIAL LOOK AT THE DATA FRAME
str(df1)
#checking the observation,resulting in 301 obs. of 23 variables and 22 columns
View(df1)
#summarizing the data set for checking number of NA's
summary(df1)


#CHECKING FOR STRUCTURAL ERRORS

#RENAMING MISSLABLED VARIABLES
library(janitor)
df1 <- clean_names(df1)
#CHECKING FOR FAULTY DATA TYPES
str(df1) # all columns are in numeric datatype except the observation columns which is char in datatype
```

```r

#ROUNDING OFF THE VALUES IN COLUMN bf_c_mratio because it has 4 decimal places
df1$bf_c_mratio <- round(df1$bf_c_mratio,digits = 2)
#DEALING WITH NA/MISPLACED/EMPTY VALUES

#replacing empty cells with NA
df1[df1==""] <- NA

#Removing the columns and rows having missing data more than 60%
#initializing a variable with the percentage value
#REMOVING ROWS WITH MORE THAN 60% NA
threshold <- 0.6 # for 60%
df1 <- df1 %>% filter(rowMeans(is.na(.))< threshold)
#REMOVING COLUMNS WITH MORE THAN 60% NA
df1 <- df1[,which(colMeans(!is.na(df1))>threshold)] #column SulphidityL_4 and AAWhiteSt_4is removed because it had more than 60% NA


summary(df1)

#Imputing the missing value

#-----------------------------------------------------------#KNN-IMPUTATION ------------------------------------------------------------------------

#The k nearest neighbours is an algorithm that is used for simple classification

#This can be very useful in making predictions about the missing values by finding the k's closest neighbours to the observation with missing data
# and then imputing them based on the non-missing values in the neighbourhood

# This KNN imputation is used in statistical data so that the NA columns are filled with appropriate nearby value,so that the calculations are made easier.

# This is a standard method used in imputing the missing values.

# The main reason of using KNN for this data set is it creates a basic mean impute then uses the resulting complete list to construct a KDTree. Then, it uses the resulting KDTree to compute nearest neighbours
#After it finds the k-NNs, it takes the weighted average of them.

#install.packages("VIM")

library(VIM)


df1 <- kNN(df1,variable = c('uczaa'),metric=NULL,k=6)
df1 <- subset(df1,select =1:21)
View(df1)


#for a particular column replacing NA with median of the column  eg:chip_rate
df1$chip_rate <- ifelse(is.na(df1$chip_rate),median(df1$chip_rate,na.rm = TRUE),df1$chip_rate)

#Applying Mean or Median for Imputing Missing Values in remaining columns[mean is used in continuous data]
```

```r
for(i in 1:ncol(df1)){
  df1[is.na(df1[,i]), i] <- mean(df1[,i], na.rm = TRUE)
}

summary(df1)
#ADDITIONAL INFORMATION
#INORDER TO CREATE A DATA SET WITH NA'S LIBRARY(missForest) provides a function called prodNA(file,noNA=0.6)


```

SAMPLE DATA SET:

Filter

	Observation	Y.Kappa	ChipRate	BF.CMratio	BlowFlow	ChipLevel4	T.upperExt.2	T.lowerExt.2	UCZAA	WhiteFlow.4	AAWhiteSt.4	AA.Wood.4	ChipMoisture.4	SteamFlow.4
1	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545	1.443	599.253	NA	16.471	46.011	67
2	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067	1.549	537.201	6.076	16.543	43.954	60
3	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260	1.600	549.611	NA	16.559	44.495	61
4	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142	1.604	623.362	6.054	16.562	45.592	68
5	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709	NA	638.672	6.110	16.677	45.512	70
6	31-05:00	20.89	14.308	94.172	1327.832	251.120	351.263	332.485	1.522	631.514	NA	16.626	45.847	71
7	31-06:00	22.65	14.100	91.887	1307.852	288.989	352.321	331.162	1.468	625.549	6.143	16.620	45.797	71
8	31-07:00	22.50	14.233	97.249	1346.900	330.325	352.687	328.894	1.480	591.827	NA	16.617	45.832	68
9	31-08:00	24.70	13.850	96.208	1334.892	362.511	352.372	327.358	1.515	553.172	6.199	16.625	45.541	64
10	31-09:00	22.60	13.875	95.720	1351.240	383.731	347.683	322.363	1.480	541.728	NA	16.624	45.695	62
11	31-10:00	24.40	14.117	85.998	1330.104	394.234	348.089	319.027	1.429	540.558	6.209	16.661	45.258	62
12	31-11:00	26.62	15.467	84.447	1334.255	386.971	349.392	321.021	1.428	531.250	NA	16.668	44.580	59
13	31-12:00	27.20	16.083	82.839	1332.331	366.855	350.094	327.439	1.486	527.893	6.190	16.669	44.541	60
14	31-13:00	24.05	16.675	77.025	1284.386	246.336	350.317	329.240	1.538	537.811	NA	16.709	44.694	58
15	31-14:00	25.40	16.425	72.924	1197.775	118.821	350.765	329.799	1.635	585.011	6.197	16.794	45.547	65
16	31-15:00	24.10	15.800	79.387	1203.375	52.241	352.311	333.573	NA	606.425	6.200	16.822	45.818	68
17	31-16:00	17.40	15.158	87.057	1218.797	6.476	350.724	330.613	1.553	628.041	NA	16.822	45.367	68
18	31-17:00	17.80	14.000	92.772	1204.489	61.783	353.562	330.481	1.532	623.209	6.200	16.824	45.441	67
19	31-18:00	20.00	12.983	100.073	1206.709	191.648	349.412	324.744	1.532	590.861	NA	16.822	45.329	64
20	31-19:00	20.30	12.058	89.523	1180.205	300.596	348.649	318.905	1.548	577.128	6.090	16.822	45.793	66
21	31-20:00	20.80	13.183	84.397	1187.186	349.159	346.360	315.365	1.509	549.357	NA	16.822	45.904	64
22	31-21:00	23.00	14.067	88.383	1191.691	374.687	346.269	318.589	1.542	515.431	6.095	16.822	46.088	62
23	31-22:00	22.70	13.483	89.578	1180.190	353.574	345.811	322.831	1.517	484.914	NA	16.822	45.541	56
24	31-23:00	21.40	12.875	92.564	1191.757	229.014	347.377	326.687	1.552	521.840	6.000	16.822	45.969	60

CODE IN R STUDIO:

```
#-----------------------------------------------------------------------------------------------------------------------

#WRITING META DATA

#USER INFROMATION : 19BDS0083 D VASANTH KUMAR

#DATA SOURCE : Kamyr digester

#OWNED BY: OpenMV.net Datasets

#Description:Pulp quality is measured by the lignin content remaining in the pulp: the Kappa number. This data set is used to understand

 # which variables in the process influence the Kappa number,and if it can be predicted accurately enough for an

  # inferential sensor application.Variables with a number at the end have been lagged by that number of hours to line up the data.

#Data shape:     301 rows and 22 columns

#TAGS FOR THE DATA SET : MULTIVARIATE,MISSING-DATA,TIME-SERIES

#-------------------------------------------------------------------------------------------------------------------------- --

#LIBRARIES USED

library(rvest)

library(dplyr)

library(tidyr)

library(utils)


#DETERMINING THE SIZE OF THE DATA SET IN MY DIRECTORY


file.info("C:/Users/vasanth kumar/Downloads/kamyr-digester.csv")$size
```

```
#READING THE DATA SET

df <-read.csv("https://openmv.net/file/kamyr-digester.csv")

#STORING THE DATA SET IN NEW VARIABLE KEEPS THE ORIGINALITY OF THE DATA SET BEFORE CLEANING

df1 <-data.frame(df)

#AN INITIAL LOOK AT THE DATA FRAME

str(df1)

#checking the observation,resulting in 301 obs. of 23 variables and 22 columns

View(df1)

#summarizing the data set for checking number of NA's

summary(df1)


#CHECKING FOR STRUCTURAL ERRORS

#RENAMING MISSLABLED VARIABLES

library(janitor)

df1 <- clean_names(df1)

#CHECKING FOR FAULTY DATA TYPES

str(df1) # all columns are in numeric datatype except the observation columns which is char in datatype


#ROUNDING OFF THE VALUES IN COLUMN bf_c_mratio because it has 4 decimal places

df1$bf_c_mratio <- round(df1$bf_c_mratio,digits = 2)

#DEALING WITH NA/MISPLACED/EMPTY VALUES


#replacing empty cells with NA

df1[df1==""] <- NA
```

#Removing the columns and rows having missing data more than 60%

#initializing a variable with the percentage value

#REMOVING ROWS WITH MORE THAN 60% NA

threshold <- 0.6 # for 60%

df1 <- df1 %>% filter(rowMeans(is.na(.))< threshold)

#REMOVING COLUMNS WITH MORE THAN 60% NA

df1 <- df1[,which(colMeans(!is.na(df1))>threshold)] #column SulphidityL_4 and AAWhiteSt_4is removed because it had more than 60% NA

summary(df1)

#Imputing the missing value

#----------#KNN-IMPUTATION ----------

#The k nearest neighbours is an algorithm that is used for simple classification

#This can be very useful in making predictions about the missing values by finding the k's closest neighbours to the observation with missing data

and then imputing them based on the non-missing values in the neighbourhood

This KNN imputation is used in statistical data so that the NA columns are filled with appropriate nearby value,so that the calculations are made easier.

This is a standard method used in imputing the missing values.

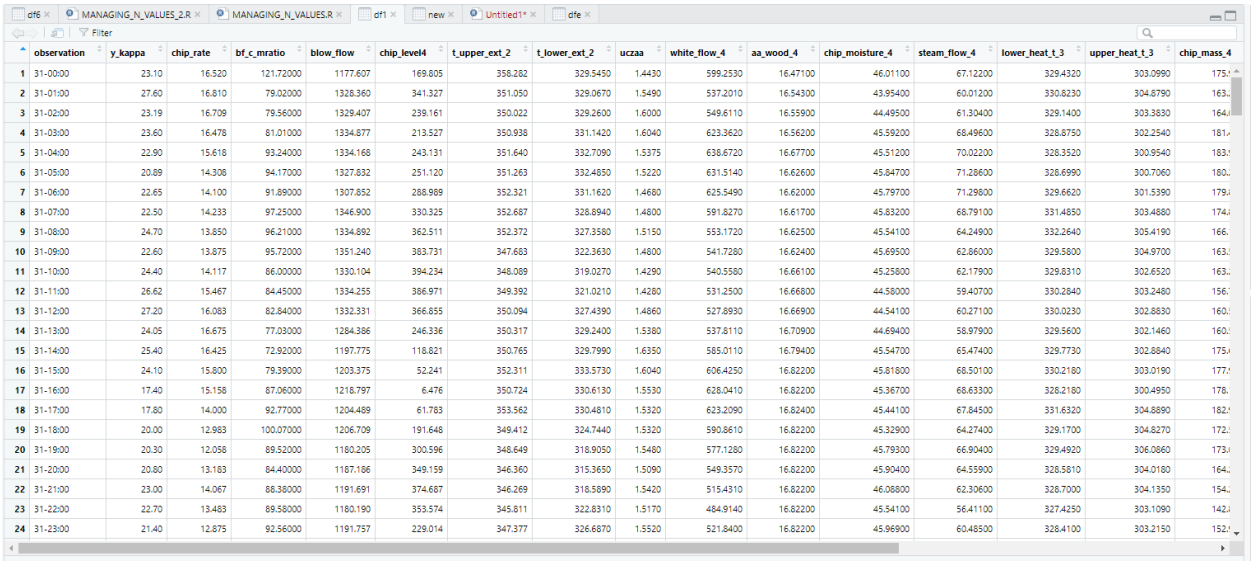
The main reason of using KNN for this data set is it creates a basic mean impute then uses the resulting complete list to construct a KDTree. Then, it uses the resulting KDTree to compute nearest neighbours (NN).

#After it finds the k-NNs, it takes the weighted average of them.

#install.packages("VIM")

```
library(VIM)

df1 <- kNN(df1,variable = c('uczaa'),metric=NULL,k=6)

df1 <- subset(df1,select =1:21)

View(df1)

#for a particular column replacing NA with median of the column  eg:chip_rate

df1$chip_rate <- ifelse(is.na(df1$chip_rate),median(df1$chip_rate,na.rm = TRUE),df1$chip_rate)

#Applying Mean or Median for Imputing Missing Values in remaining columns[mean is used in continuous data]

for(i in 1:ncol(df1)){

  df1[is.na(df1[,i]), i] <- mean(df1[,i], na.rm = TRUE)

}

summary(df1)

#ADDITIONAL INFORMATION

#INORDER TO CREATE A DATA SET WITH NA'S LIBRARY(missForest) provides a function called prodNA(file,noNA=0.6)
```

OUTPUT AFTER DEALING WITH NA’s:



	observation	y_kappa	chip_rate	bf_c_mratio	blow_flow	chip_level4	t_upper_ext_2	t_lower_ext_2	uczaa	white_flow_4	aa_wood_4	chip_moisture_4	steam_flow_4	lower_heat_t_3	upper_heat_t_3	chip_mass_4
1	31-00:00	23.10	16.520	121.72000	1177.607	169.805	358.282	329.5450	1.4430	599.2530	16.47100	46.01100	67.12200	329.4320	303.0990	175.
2	31-01:00	27.60	16.810	79.02000	1328.360	341.327	351.050	329.0670	1.5490	537.2010	16.54300	43.95400	60.01200	330.8230	304.8790	163.
3	31-02:00	23.19	16.709	79.56000	1329.407	239.161	350.022	329.2600	1.6000	549.6110	16.55900	44.49500	61.30400	329.1400	303.3830	164.
4	31-03:00	23.60	16.478	81.01000	1334.877	213.527	350.938	331.1420	1.6040	623.3620	16.56200	45.59200	68.49600	328.8750	302.2540	181.
5	31-04:00	22.90	15.618	93.24000	1334.168	243.131	351.640	332.7090	1.5375	638.6720	16.67700	45.51200	70.02200	328.3520	300.9540	183.
6	31-05:00	20.89	14.308	94.17000	1327.832	251.120	351.263	332.4850	1.5220	631.5140	16.62600	45.84700	71.28600	328.6990	300.7060	180.
7	31-06:00	22.65	14.100	91.89000	1307.652	288.989	352.321	331.1620	1.4680	625.5490	16.62000	45.79700	71.29800	329.6620	301.5390	179.
8	31-07:00	22.50	14.233	97.25000	1346.900	330.325	352.687	328.8940	1.4800	591.8270	16.61700	45.83200	68.79100	331.4850	303.4880	174.
9	31-08:00	24.70	13.850	96.21000	1334.892	362.511	352.372	327.3580	1.5150	553.1720	16.62500	45.54100	64.24900	332.2640	305.4190	166.
10	31-09:00	22.60	13.875	95.72000	1351.240	383.731	347.683	322.3630	1.4800	541.7280	16.62400	45.69500	62.86000	329.5800	304.9700	163.
11	31-10:00	24.40	14.117	86.00000	1330.104	394.234	348.089	319.0270	1.4290	540.5580	16.66100	45.25800	62.17900	329.8310	302.6520	163.
12	31-11:00	26.62	15.467	84.45000	1334.255	386.971	349.392	321.0210	1.4280	531.2500	16.66800	44.58000	59.40700	330.2840	303.2480	156.
13	31-12:00	27.20	16.083	82.84000	1332.331	366.855	350.094	327.4390	1.4860	527.8930	16.66900	44.54100	60.27100	330.0230	302.8830	160.
14	31-13:00	24.05	16.675	77.03000	1284.386	246.336	350.317	329.2400	1.5380	537.8110	16.70900	44.69400	58.97900	329.5600	302.1460	160.
15	31-14:00	25.40	16.425	72.92000	1197.775	118.821	350.765	329.7990	1.6350	585.0110	16.79400	45.54700	65.47400	329.7730	302.8840	175.
16	31-15:00	24.10	15.800	79.39000	1203.375	52.241	352.311	333.5730	1.6040	606.4250	16.82200	45.81800	68.50100	330.2180	303.0190	177.
17	31-16:00	17.40	15.158	87.06000	1218.797	6.476	350.724	330.6130	1.5530	628.0410	16.82200	45.36700	68.63300	328.2180	300.4950	178.
18	31-17:00	17.80	14.000	92.77000	1204.489	61.783	353.562	330.4810	1.5320	623.2090	16.82400	45.44100	67.84500	331.6320	304.8890	182.
19	31-18:00	20.00	12.983	100.07000	1206.709	191.648	349.412	324.7440	1.5320	590.8610	16.82200	45.32900	64.27400	329.1700	304.8270	172.
20	31-19:00	20.30	12.058	89.52000	1180.205	300.596	348.649	318.9050	1.5480	577.1280	16.82200	45.79300	66.90400	329.4920	306.0860	173.
21	31-20:00	20.80	13.183	84.40000	1187.186	349.159	346.360	315.3650	1.5090	549.3570	16.82200	45.90400	64.55900	328.5810	304.0180	164.
22	31-21:00	23.00	14.067	88.38000	1191.691	374.687	346.269	318.5890	1.5420	515.4310	16.82200	46.08800	62.30600	328.7000	304.1350	154.
23	31-22:00	22.70	13.483	89.58000	1180.190	353.574	345.811	322.8310	1.5170	484.9140	16.82200	45.54100	56.41100	327.4250	303.1090	142.
24	31-23:00	21.40	12.875	92.56000	1191.757	229.014	347.377	326.6870	1.5520	521.8400	16.82200	45.96900	60.48500	328.4100	303.2150	152.