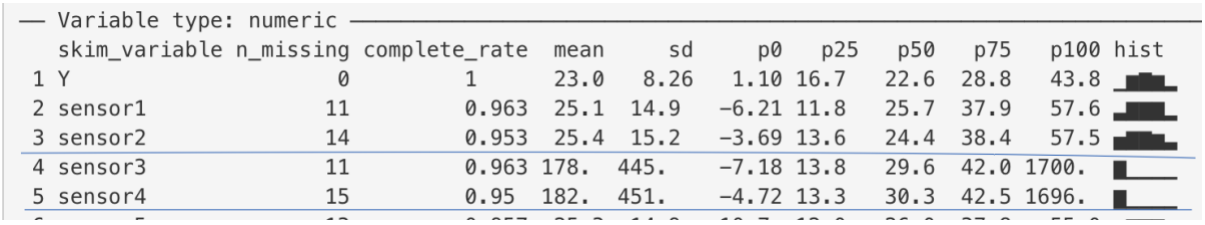
Exploratory Data Analysis Report

Name: Shi Chen
E-mail: yuejianyingmm@icloud.com

This document is discussing the interesting features I found in the Ass1Data. The Ass1Data is a .csv file, which has 300 rows(observations) and 44 columns(variables). There is no background information about this dataset. We can only find some novelties from the statistics summary and data morphology description. They will help us to understand the basic information of the data set and suggests some possible relationships between variables, as well as some abnormal data that can be further explored. In this assignment, we use Shinny R to show the data description. And I will discuss what I found in summary, missing value figure, "Corrgram" matrix, mixed pairs groups, boxplots chart, rising value chart and mosaic chart. The public link for this assignment is https://sch405.shinyapps.io/Assignment1/ .

From the "Summary", in the value distribution of numeric variables part. I found the value distribution of sensor 3, sensor 4, sensor 13, sensor 17, sensor 22, sensor 24 and sensor 27 is interesting(shown in figure 1), which may have outstanding outliers, while other values excluding outliers have small standard deviation. As we can see from the P100 values of these variables, which are around 1690, they have huge difference from their median values and other variables' p100 values. Do the outliers reveal some recording errors? Are they concurrency in the same records? Actually, I found some clues which will discuss later.



```
── Variable type: numeric ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate  mean     sd     p0   p25   p50   p75   p100 hist
1 Y                     0             1  23.0   8.26   1.10 16.7   22.6  28.8   43.8
2 sensor1              11         0.963  25.1   14.9  -6.21 11.8   25.7  37.9   57.6
3 sensor2              14         0.953  25.4   15.2  -3.69 13.6   24.4  38.4   57.5
4 sensor3              11         0.963 178.   445.   -7.18 13.8   29.6  42.0 1700.
5 sensor4              15         0.95  182.   451.   -4.72 13.3   30.3  42.5 1696.
```

(Figure1: interesting distribution of sensor 3 and sensor 4)

From the "Missing Value", there is no missing value of the variable set {Y, ID, Author, Date, Priority}. The variable sensor 7 holds 22% missing value, that is interesting. What is the meaning of this variable? Why does it have so many missing values? Is it related to some particular time or a particular author?

From the "Corrgram", we see that the whole numeric variables are grouped into three clusters by using "Spearman" and "Kendall" methods. But if we change the method into "Pearson", it will generate 4 highly correlated variable clusters. By analyzing the detail, I found that the "Pearson" method groups { sensor 3, sensor 4, sensor 13, sensor 17, sensor 22, sensor 24, sensor 27 } together as an extra cluster, which I mentioned before, they contain a lot of outstanding outliers. Despite that, however, the variables in each cluster under different methods almost have not too much difference.

In the "mixed pairs" part, I decided to use the numeric clusters grouped in "corrgram" and added a little modify as the mixed pairs groups, which shown in "Mixed Pairs set1", "Mixed Pairs set2" and "Mixed Pairs set3" separately. Then I added another group which contains factor variables and variable Y, which shown in "Mixed Pairs set4". From "Mixed Pairs set1", "Mixed Pairs set2" and "Mixed Pairs set3", I found three kinds of interesting information which confirmed again in the plots graph. They are missing value variable, and the variables which have outstanding outliers, and some linear relationship variables set, and the variable which has no correlation to others in the group. Their spots are special and the examples are shown below(figure 2).

(Figure 2: different type of interesting plots)

Here is some guess from the "mixed pairs":

(1) If I want to do a "formula-based" model analysis for the outcome Y, I will start from set1(most of which seems correlates with Y) and drop the sensor 11(which does not correlate with Y).

(2) Considering the linear relationship between the predictors {sensor1, sensor 2, sensor 5, sensor 6, sensor 7, sensor 8, sensor 9, sensor 10 }, but multicollinearity in predictors is not a good phenomenon, it is better only choose one of them which has the highest correlation coefficient(sensor 9).

(figure 3:sensor 9 has the highest correlation coefficient with Y in these group)

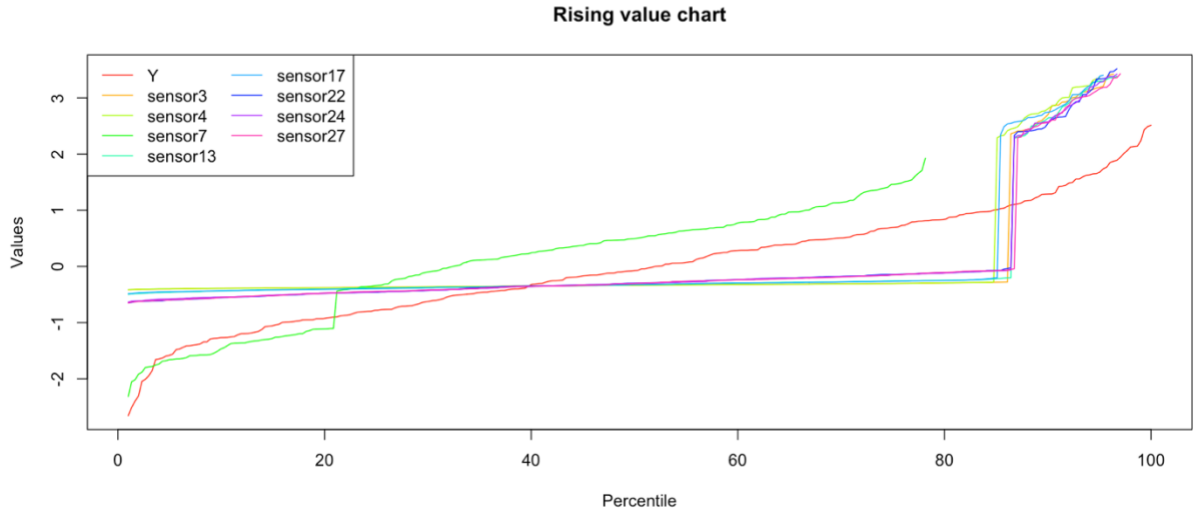
(3) The linear relationship variable set may help us to predict the missing values of sensor 9.

From "Boxplot & Outliers" part, we got confirm information from the figure that we already know from the summary. The outlier plots massive part are the variables which hold outstanding outliers. Without tick the selection "show standardized", based on the height, we can see the boxes can be roughly divided into three groups. But when we tick the selection "show standardized", the groups information is hidden, but the outliers are much easy to find out. Their tiny height is very conspicuous. Another interesting part is about the outliers, from the vertical axis, we can see the distribution of the outliers in different variables are almost in a similar range([1100, 1700]). Are they the features from the same observations, are they recorded by the same author?

From "Rising value" part, by analyzing the shape of the whole bunch of the numeric variables, I picked the most interesting group, {Y, sensor3, sensor4, sensor7, sensor13, sensor17, sensor22, sensor24, sensor27}, as default option. As shown below(figure 4), the rising line of Y is a standard continuous rise. In compare, the larger value part at the end of the increasing value curve of sensor 7 seems to be missing, and the value around [20, 23] percentile has a discontinuous high jump, which suggests some issues. At the same time, as mentioned before, the notable outliers dataset {sensor3, sensor4, sensor13, sensor17, sensor22, sensor24, sensor27}, shows similar large-span high jumps in the shape of the rising value curve, and at the tail (the outliers part) which suggests the larger value, the fluctuates of the curve are more frequently and not smooth, which is different from the shape in the front section. Is that possible that the outliers are fake and added by someone else?

Select variable set:

Y sensor3 sensor4 sensor7 sensor13
sensor17 sensor22 sensor24 sensor27



(figure4: The shape of interesting rising value dataset)

From the "Mosaic" part, I tried many combinations, some shows slight blue and the red color but no dark red been found. It suggests that there is no significant pattern or uncommon cases in this category variable dataset. Maybe adding fake data to give some highlights in category variables is not as easy as it to numeric variables.

To solve the previous doubts, I use excel to help me find some more information.

(1) Does the missing value of sensor7 occurre in the records from the same author, who is careless? The answer is not really. According to the figure 5, we can see that the missing value of sensor7 occurred in the author HH's records are more than the other two, but the total amount of he/she's work is also far beyond the other two. After comparing relative value and absolute value, we cannot say the missing value is related to a particular author.

Table 1: The records of Missing value of sensor 7 for each author

Row Labels	count:Missin	count:record	Persantage
HH	40	180	22%
KG	17	70	24%
KL	6	19	32%
XX	3	31	10%
Grand Total	**66**	**300**	**22%**

(figure 5: The records of missing value of sensor 7 for each author)

(2) Does the outliers of these seven numerical variables from the same records and same author? The answer is yes. From the “Boxplot & Outliers” part and summary part, we got some essential features of these outliers, such which variables are they belong, and their value range. Using this information, it is easy for us to locate the records, see below(figure 6). The outliers occur in these 31 records all from the author XX, the value range of the outliers is from [1200, 17500], and they occur in different periods(the dates of these records are random). If we want to know more about the outliers, maybe we can ask XX.

Seq	Y	ID	Auth	Date	sensor3	sensor4	sensor13	sensor1	sensor22	sensor24
12	12.3745685	D282	XX	2006/5/28	1699.997022	1631.033775	1528.9991	1362.656	1396.95948	1513.48205
13	26.1736659	D284	XX	2006/6/11	1695.432609	1511.703141	1369.9947	1567.35	1298.86672	1267.54317
14	14.841968	D277	XX	2006/4/23	1675.465385	1469.720364	1589.1909	1690.702	1679.14914	1452.73839
15	25.390921	D280	XX	2006/5/13	1604.607668	1625.34506	1527.0879	1376.244	1233.27588	1321.95279
16	34.7857662	D271	XX	2006/3/11	1596.636501	1632.633234	1306.2361	1379.103	1203.36012	1413.44121
17	22.3520475	D289	XX	2006/7/16	1591.516925	1431.13788	1371.0891	1291.265	1676.33267	1314.37098
18	11.9473629	D272	XX	2006/3/19	1579.590444	1675.252607	1604.8691	1684.734	1340.41236	1441.37464
19	23.098855	D283	XX	2006/6/4	1575.2061	1474.144625	1323.1376	1312.918	1237.13409	1321.53528
20	11.5306387	D298	XX	2006/9/17	1561.616908	1426.185135	1653.2506	1330.784	1264.96729	1622.2944
21	21.7385436	D276	XX	2006/4/16	1559.624003	1284.606207	1225.2806	1638.526	1237.74102	1203.29472
22	27.1770599	D295	XX	2006/8/26	1542.44754	1615.230296	1218.5761	1317.362	1625.97397	1312.98957
23	17.2666363	D290	XX	2006/7/23	1518.778752	1303.097331	1644.5689	1366.261	1630.5762	1371.5606
24	9.95373082	D296	XX	2006/9/2	1515.092715	1449.986388	1414.2817	1540.103	1529.07295	1656.82855
25	29.5839409	D270	XX	2006/3/4	1493.038572	1409.932531	1685.1086	1334.437	1334.62059	1517.47783
26	16.6959454	D291	XX	2006/7/29	1481.435262	1407.018106	1335.6052	1404.109	1341.53115	1581.33957
27	21.4448767	D292	XX	2006/8/6	1480.099734	1539.039988	1683.3937	1423.278	1454.13834	1664.70987
28	26.9376148	D281	XX	2006/5/20	1457.285026	1230.125169	1306.489	1439.519	1552.04597	1397.53324
29	28.6249697	D287	XX	2006/7/2	1455.271237	1538.460695	1656.9877	1541.45	1453.75784	1664.81977
30	27.3458035	D300	XX	2006/10/1	1451.847286	1334.439343	1570.6631	1410.469	1321.99711	1467.3077
31	18.015164	D294	XX	2006/8/19	1385.033634	1686.896708	1467.3124	1362.27	1507.53439	1682.45983
32	15.1828283	D286	XX	2006/6/24	1358.415603	1373.611098	1508.558	1608.482	1622.21836	1372.52694
33	19.2467491	D288	XX	2006/7/9	1333.993527	1393.238516	1486.3528	1401.455	1231.82864	1599.05233
34	17.1691318	D274	XX	2006/4/1	1324.446924	1404.166523	1694.3413	1541.156	1698.58423	1231.48771
35	16.6695823	D293	XX	2006/8/12	1302.762506	1621.28891	1229.3355	1229.231	1230.90792	1219.97238
36	17.0970776	D297	XX	2006/9/9	1268.464242	1234.621737	1378.4593	1343.004	1251.26898	1485.19653
37	13.3694941	D285	XX	2006/6/18	1261.508008	1622.893876	1454.1861	1453.881	1251.94287	1248.18273
38	23.5088861	D278	XX	2006/4/29	1257.72322	1291.473886	1671.605	1326.913	1243.63008	1218.93364
39	20.7086576	D299	XX	2006/9/24	1251.443196	1696.18919	1208.0616	1640.995	1507.94069	1350.35883
40	41.5203613	D275	XX	2006/4/8	1250.902552	1271.485366	1349.7223	1593.186	1624.01963	1625.99583
41	22.0338446	D273	XX	2006/3/26	1238.39905	1534.599397	1657.5474	1467.653	1344.83654	1590.11606
42	31.128121	D279	XX	2006/5/7	1229.116467	1215.21517	1251.9296	1492.974	1434.10608	1313.04098

(figure 6: The records of outliers of the seven variables)