Modern application for Data Science Final Exam

Sinney Chan

Lipscomb University

12th December 2016

Abstract

This paper includes two parts of the exam. The first part of the paper reports the results of a set of regression data by using two different software, which I chose R and Tableau. The report includes the scatter plots that shows linear regression between two variables and the summary of the plots. The second part of the paper is to exam tumor gene expression data by running clustering tools to group samples into five cluster and compare the output with the another set of tumor subtypes data. After several attempts of running the data sets in different software, I got the best result by using Tableau's analytic clustering tool.

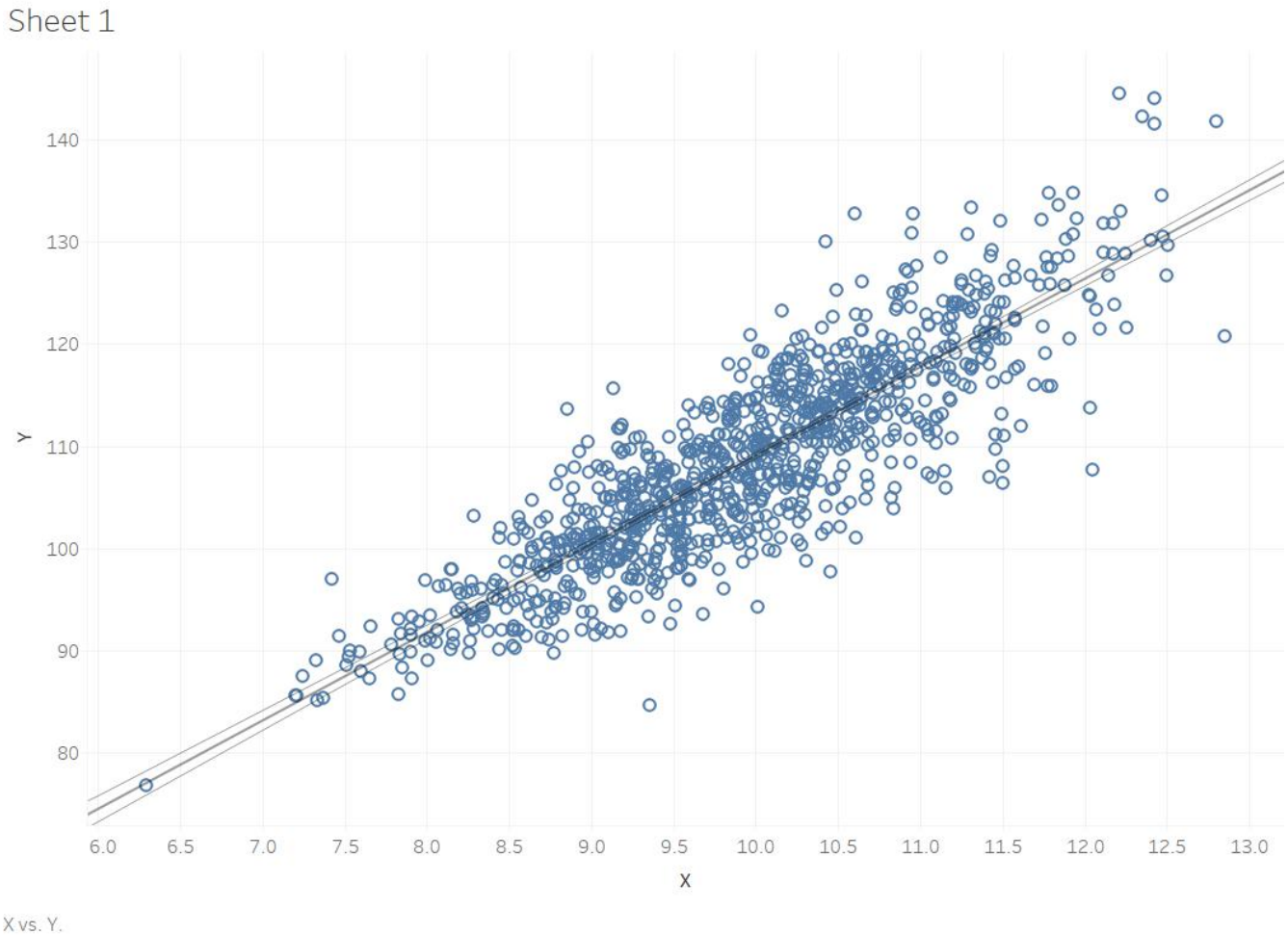
Part One

Part one of this paper is to identify the relationship of two variables of the given data set. The first graph is generated by using Tableau. From observing the plot, we can see that there is a positive relationship between X and Y. Moreover, from the description of trend analysis, we know that the correlation value r^2 = 0.748, which indicates a strong and positive relationship between X and Y.

The second graph is a scatter plot generated by using R. It shows a fairly strong and reasonable linear relationship between two variables. We also know that r^2 = 0.7482, which indicates a strong positive relationship. The estimates for the model intercept is 0.54 and the coefficient measuring the slope of the relationship with x is 0.087 and information about standard errors of these estimates is also provided in the Coefficients table.

Both graphs suggest that y increase linearly with x. As well as having corresponding correlation value r^2 = 0.748. Both graphs have intercept of 22.64. We can say that both graphs are corresponding to each other.

Graph 1:



Trend Lines Model

A linear trend model is computed for Y given X. The model may be significant at p <= 0.05.

Model formula: (X + intercept)
Number of modeled observations: 1000
Number of filtered observations: 0
Model degrees of freedom: 2
Residual degrees of freedom (DF): 998
SSE (sum squared error): 26314.8
MSE (mean squared error): 26.3675
R-Squared: 0.748482
Standard error: 5.13493
p-value (significance): < 0.0001

Individual trend lines:

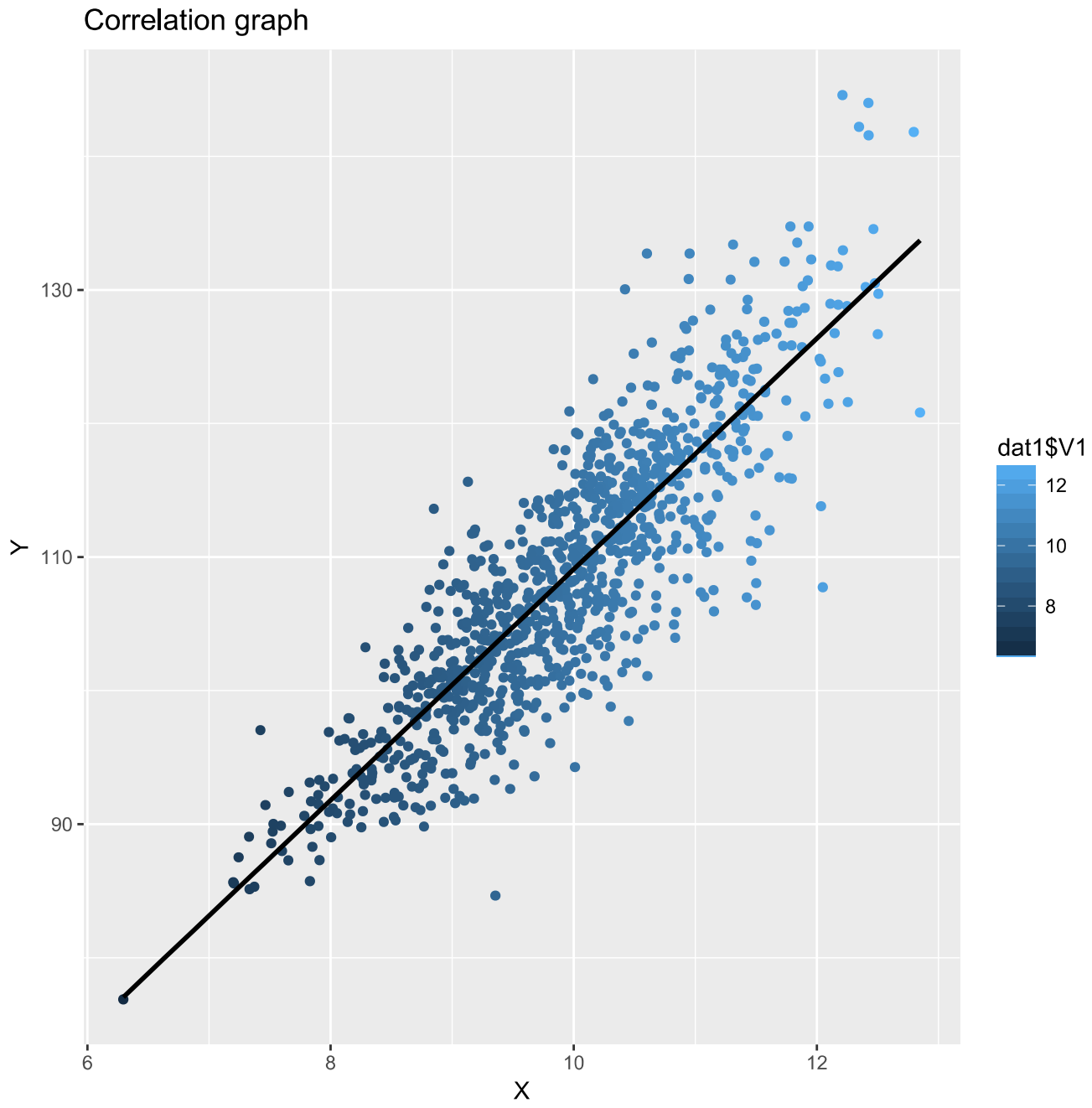
Panes		Line		Coefficients				
Row	Column	p-value	DF	Term	Value	StdErr	t-value	p-value
Y	X	< 0.0001	998	X	8.64372	0.15861	54.4968	< 0.0001
				intercept	22.6407	1.58522	14.2823	< 0.0001

Graph 2:

The following part is result generated by using R.

```
getwd()
setwd("C:/Users/sing_/Desktop/2016 Fall/ModernAppDS/")
library(ggplot2)
dat1 <- read.csv("final/regressiondata.csv", header=FALSE)
pc1 <- ggplot(dat1, aes(x=dat1$V1,y=dat1$V2, color=dat1$V1))+geom_point()

(curveplot <- pc1 +
  geom_smooth(aes(group = 1),
              method = "lm",
              formula = y ~ x,
              se = FALSE,
              color = "black")) +
  ggtitle ("Correlation graph")+
  labs(x="X",y="Y")
  geom_point()

summary(lm(V2 ~ V1, data = dat1))


Call:
lm(formula = V2 ~ V1, data = dat1)

Residuals:
     Min       1Q   Median       3Q      Max
-19.0448  -3.0840   0.1787   3.1714  18.4397
```

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  22.6407      1.5852    14.28    <2e-16
V1            8.6437      0.1586    54.50    <2e-16

(Intercept) ***
V1          ***
---
Signif. codes:
0 ?***?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 5.135 on 998 degrees of freedom
Multiple R-squared:  0.7485,    Adjusted R-squared:  0.7482
F-statistic:  2970 on 1 and 998 DF,  p-value: < 2.2e-16
```

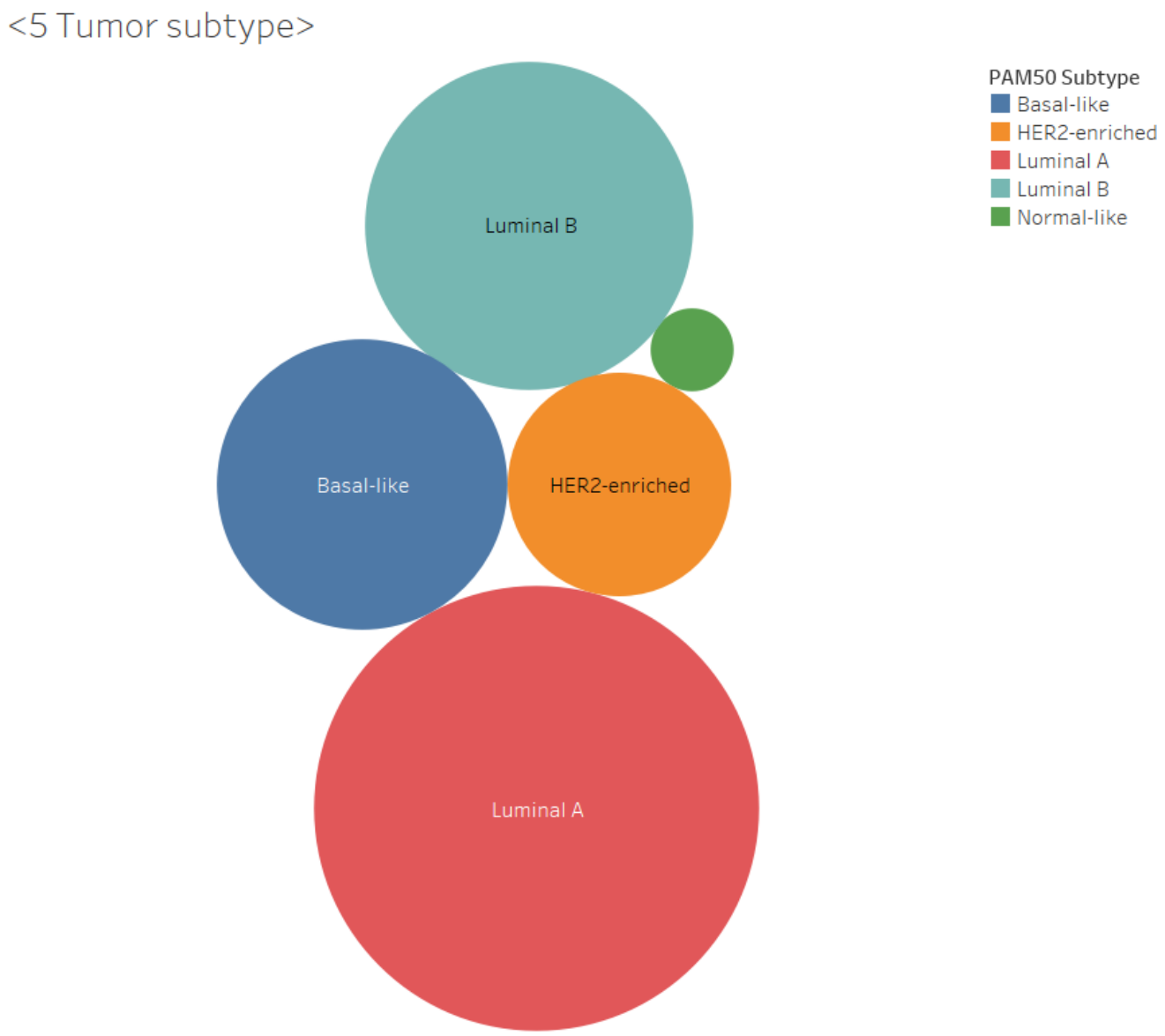


Part Two

Part two is to run data sets through clustering tools and compare the results of both data sets, which are clustering-data and tumor-subtype.

I used Tableau to analysis the data sets, for the clustering data, I set the sum of the number of records as the size and PAM50 Subtype as the label and color. Then run the analytic cluster model through the data, Tableau will then generate the following graph of the clustering data set, as well as the description of clusters. We can see that Luminal A is the biggest cluster, it has 230 number of records; Normal-like is the smallest cluster and it only have 8 number of records. And for the tumor-type data, I used the same method to cluster the data, however before running the data set through Tableau, I had to transpose the tumor-subtype data and delete the first column of the data in R, so that the data set would be ready to be analyzed in Tableau. I repeated the steps of analyzing the clustering data in Tableau, and got the result in Graph 4. We can see that cluster one has the most dots, and cluster 4 has the least amount of dots. Each dot represent one gene expression. In table 1, we can see how graph 3 and 4 correspond to each others. The biggest difference is that in graph 3, 24 percent of the number of records is Luminal B, however in graph 4, 34% of the records is in cluster 3. And in both graphs, there are 8 gene expressions are in the Normal-like cluster.

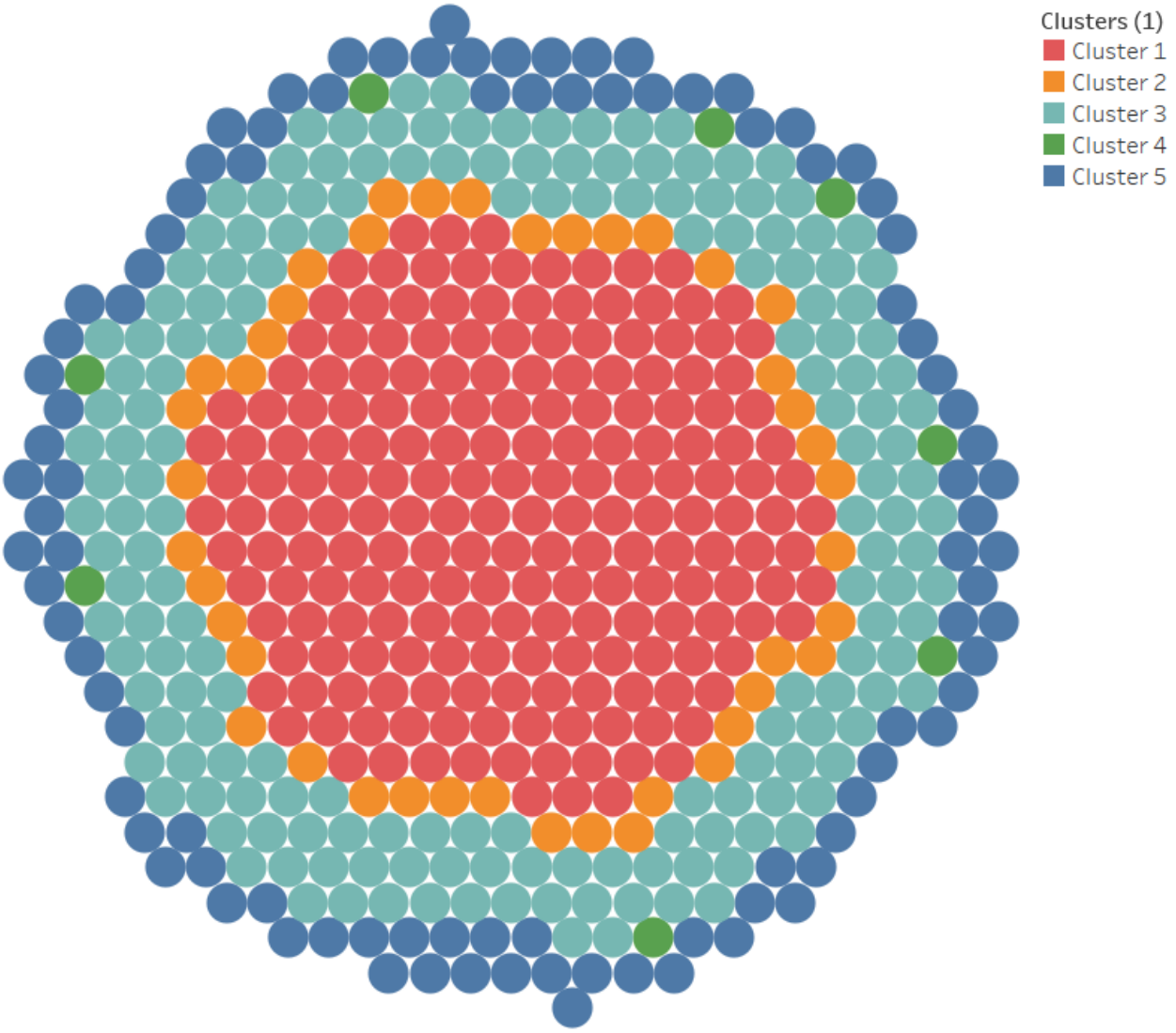
Graph 3:



PAM50 Subtype	Number of Records
Normal-like	8
Luminal B	125
Luminal A	230
HER2-enriched	58
Basal-like	98

Graph 4:

<TumorDataClustering
>



Clusters (1) and F1. Color shows details about Clusters (1). Size shows sum of Number of Records. The marks are labeled by Clusters (1) and F1.

Inputs for Clustering

Variables: Avg. F37
Level of Detail: F1
Scaling: Normalized

Summary Diagnostics

Number of Clusters: 5
Number of Points: 519
Between-group Sum of Squares: 9.369
Within-group Sum of Squares: 1.01
Total Sum of Squares: 10.38

		Centers
Clusters	**Number of Items**	**Avg. F37**
Cluster 1	198	-0.075
Cluster 2	42	2.4
Cluster 3	175	1.0
Cluster 4	8	4.5
Cluster 5	96	-1.3
Not Clustered	0	

Table 1:

Normal-like	8	2%			
Luminal B	125	24%			
Luminal A	230	44%			
HER2-enriched	58	11%			
Basal-like	98	19%			

R code:

```
getwd()
setwd("C:/Users/sing_/Desktop/2016 Fall/ModernAppDS/")
dat <- read.table("final/tumor-subtype.txt",header = TRUE, sep =
"\t")
data2 = t(dat[-c(1)])

library(xlsx)
write.xlsx(data2, "C:/Users/sing_/Desktop/2016
Fall/ModernAppDS/final/data2.xlsx")
```