Introduction to R, module05

Steve Simon

Created 2020-04-02

Introduction

```
suppressMessages(
  suppressWarnings(
    library(tidyverse)))
R.version.string
## [1] "R version 4.1.1 (2021-08-10)"
Sys.Date()
## [1] "2022-05-02"
```

Like the earlier presentations, this Powerpoint file was created using Rmarkdown.

How do you characterize relationships?

- Between two continuous variables
  - Correlations and scatterplots
- Between two categorical variables
  - Crosstabulations
- Between a continuous variable and a categorical variable
  - Boxplots

In an earlier module, you saw datasets that had mostly continuous variables. If you wanted to examine the relationship between two continuous variables, you would look at correlations and scatterplots.

Then in a different module, you saw datasets that had mostly categorical variables. If you wanted to examine the relationship between two categorical variables, you would look at crosstabulations.

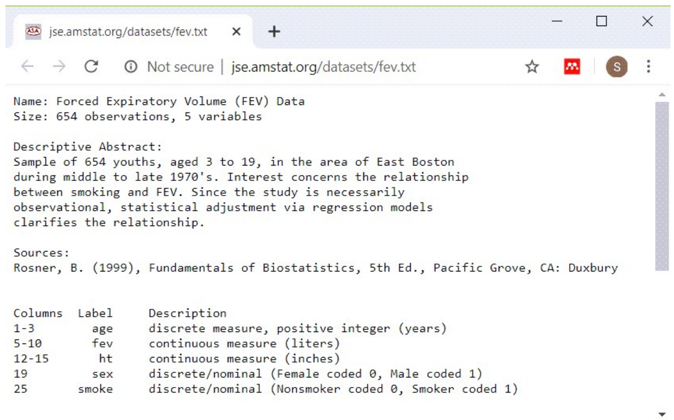
In this module, you will see datasets that have a mix of continuous and categorical variables. If you want to examine the relationship between a continuous variable and a categorical variable, you would use a boxplot.

FEV data

- FEV dataset
  - http://www.amstat.org/publications/jse/datasets/fev.dat.txt
- FEV data dictionary
  - http://ww2.amstat.org/publications/jse/datasets/fev.txt

The first data set looks at pulmonary function in a group of children. The acronym FEV stands for Forced Expiratory Volume and represents how air you can blow out of your lungs.

FEV data dictionary



Name: Forced Expiratory Volume (FEV) Data
Size: 654 observations, 5 variables

Descriptive Abstract:
Sample of 654 youths, aged 3 to 19, in the area of East Boston during middle to late 1970's. Interest concerns the relationship between smoking and FEV. Since the study is necessarily observational, statistical adjustment via regression models clarifies the relationship.

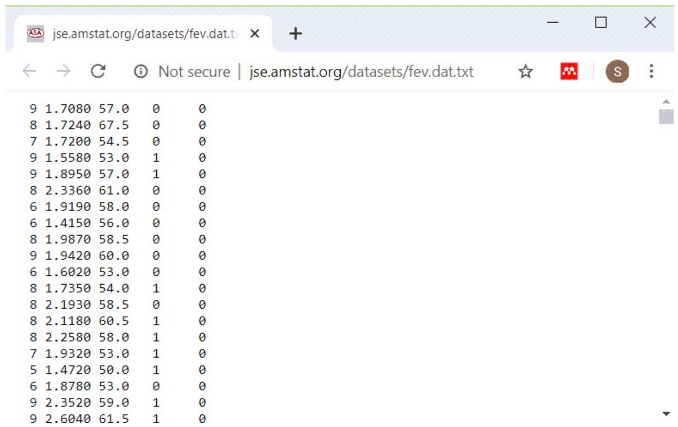
Sources:
Rosner, B. (1999), Fundamentals of Biostatistics, 5th Ed., Pacific Grove, CA: Duxbury

Columns	Label	Description
1-3	age	discrete measure, positive integer (years)
5-10	fev	continuous measure (liters)
12-15	ht	continuous measure (inches)
19	sex	discrete/nominal (Female coded 0, Male coded 1)
25	smoke	discrete/nominal (Nonsmoker coded 0, Smoker coded 1)

Screenshot of data dictionary page

This dataset has 654 rows and 5 variables: age (in years), fev (in liters), height (in inches), sex, and smoking status. Both sex and smoking status are categorical and use number codes.

Peek at FEV dataset



9	1.7080	57.0	0	0
8	1.7240	67.5	0	0
7	1.7200	54.5	0	0
9	1.5580	53.0	1	0
9	1.8950	57.0	1	0
8	2.3360	61.0	0	0
6	1.9190	58.0	0	0
6	1.4150	56.0	0	0
8	1.9870	58.5	0	0
9	1.9420	60.0	0	0
6	1.6020	53.0	0	0
8	1.7350	54.0	1	0
8	2.1930	58.5	0	0
8	2.1180	60.5	1	0
8	2.2580	58.0	1	0
7	1.9320	53.0	1	0
5	1.4720	50.0	1	0
6	1.8780	53.0	0	0
9	2.3520	59.0	1	0
9	2.6040	61.5	1	0

Listing of first few lines of data

This is a listing of the first few rows. It could be a tab delimited file or a fixed width file. If you look carefully at the data, you will see that there are blanks and no tabs. So this is a file that you can read in most easily using a fixed width format.

read in the FEV data set, code

```
fn <- "../data/fev.txt"
fev <- read_fwf(file=fn,
  col_types="nnnnn",
  col_positions=fwf_cols(
    age=3,
    fev=7,
    ht=5,
    sex=4,
    smoke=6))
```

If you count carefully, you will see that the first three columns represent the first variable, you need seven more columns for the second variable, and so forth.

read in the FEV data set, glimpse

```
glimpse(fev)
## Rows: 654
## Columns: 5
## $ age   <dbl> 9, 8, 7, 9, 9, 8, 6, 6, ~
## $ fev   <dbl> 1.708, 1.724, 1.720, 1.5~
## $ ht    <dbl> 57.0, 67.5, 54.5, 53.0, ~
## $ sex   <dbl> 0, 0, 0, 1, 1, 0, 0, 0, ~
## $ smoke <dbl> 0, 0, 0, 0, 0, 0, 0, 0, ~
```

Here is the structure of the data frame.

Summary for continuous variables: age

```
mean(fev$age)
## [1] 9.931193
sd(fev$age)
## [1] 2.953935
range(fev$age)
## [1]  3 19
```

This is clearly a pediatric population. You could use the summary function here, as it provides the mean, median, quartiles and minimum/maximum, but unfortunately, it does not include the standard deviation.

Summary for continuous variables: fev

```
mean(fev$fev)
## [1] 2.63678
sd(fev$fev)
## [1] 0.8670591
range(fev$fev)
## [1] 0.791 5.793
```

I am not an expert on FEV, but these values seem reasonable.

Summary for continuous variables: ht

```
mean(fev$ht)
## [1] 61.14358
sd(fev$ht)
## [1] 5.703513
range(fev$ht)
## [1] 46 74
```

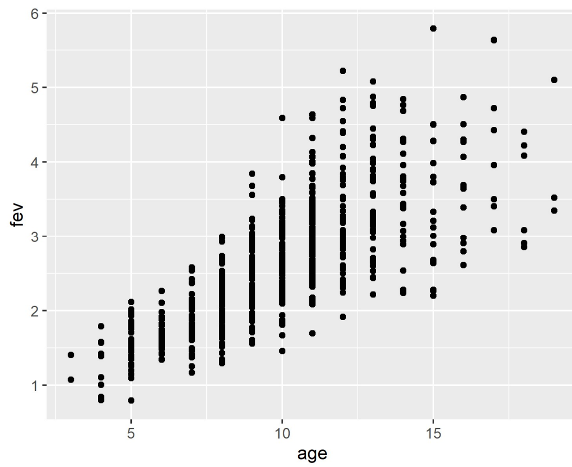
Again, these seem to be reasonable values.

Scatterplot

```
plot1 <- ggplot(fev, aes(x=age, y=fev)) +
  geom_point()
ggsave(
  "../images/age-by-fev.png",
  plot1)
## Saving 5 x 4 in image
```

Recall that you use a scatterplot to examine the relationship between two continuous variables.

Scatterplot



Here is the plot that is produced by this code.

Create factors

```
fev$smoke_factor <- factor(
  fev$smoke,
  levels=0:1,
  labels=c("nonsmoker", "smoker"))
fev$sex_factor <- factor(
  fev$sex,
  levels=0:1,
  labels=c("female", "male"))
```

When you have number codes for categorical data, it is always a good idea to create factors. Remember, though, that you should not create factors until most of the data management tasks (e.g., recoding) is done.

FEV frequency tables

```
table(fev$smoke_factor, useNA="always")
##
## nonsmoker    smoker      <NA>
##       589        65         0
table(fev$sex_factor, useNA="always")
##
## female   male   <NA>
##    318    336      0
```

The two categorical variables have no missing values.

Crosstabs

```
crosstab <-
  table(fev$sex_factor,fev$smoke_factor)
prop_table <- prop.table(crosstab,
margin=1)
pct_table <- round(100*prop_table)
```

Also recall that you use a crosstabulation to examine the relationship between two categorical variables.

The general advice, which works more than 90% of the time is to place your outcome variable as the columns and ask for row percents. Recall that margin=1 provides row percents.

The females smoke more often than the males, 12% versus 8%.

Break #1

- What have you learned
  - Review analysis of continuous variables
  - Review analysis of categorical variables
- What's next
  - Analysis of a mix: continuous and categorical

Let's take a short break here. What you've seen so far is review of methods that you use when examining relationships between two continuous variables (scatterplots) and methods that you use when examining relationships between two categorical variables (crosstabulations).

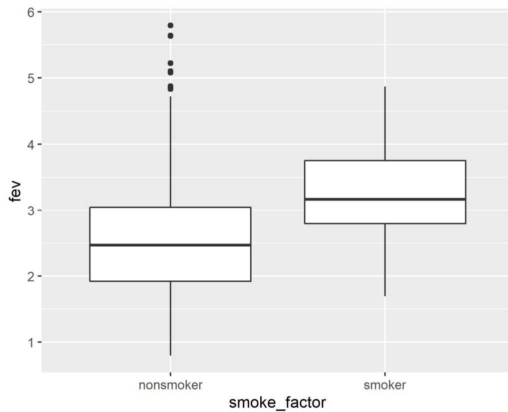
Coming up next is how you examine relationships between a continuous variable and a categorical variable.

Boxplots

```
plot2 <-
  ggplot(fev,
    aes(x=smoke_factor, y=fev)) +
      geom_boxplot()
ggsave("../images/smoke-by-fev.png")
## Saving 5 x 4 in image
```

When you want to look at a relationship between a categorical variable and a continuous variable, you should use a boxplot.

Boxplots



Boxplots comparing FEV values for smokers and non-smokers

Here is the boxplot. The results are very odd. Smokers tend to have higher FEV values than non-smokers.

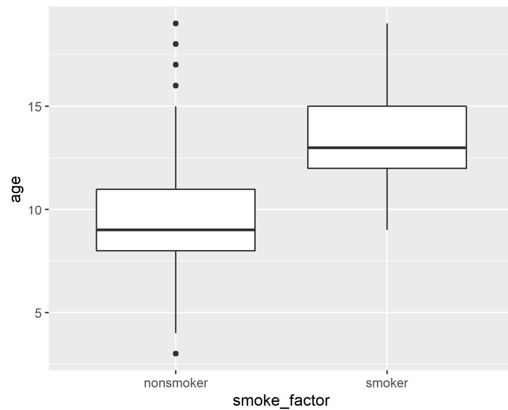
You can get a hint as to why smokers might have higher fev values than non-smokers by looking at how age and smoking status are related.

Boxplots

```
plot3 <-
  ggplot(fev, aes(smoke_factor, age)) +
    geom_boxplot()
ggsave(
  "../images/smoke-by-age.png",
  plot3)
## Saving 5 x 4 in image
```

This is the code to draw a boxplot of the ages for smokers and non-smokers

Boxplots



Boxplots comparing ages for smokers and non-smokers

Here’s what’s happening. Older kids are more likely to be smokers, and older kids have bigger and more mature lungs. This a classic case of confounding.

Break #2

- What have you learned?
  - Boxplots
- What's coming up next?
  - Group means and standard deviations

Time for another break. You saw how to use boxplots to visually compare a continuous variable across different levels of a categorical variable.

Coming up next is the calculation of group means and standard deviations.

Group means, code

```
fev_means <-
  by(
    fev$fev,
    fev$smoke_factor,
    mean,
    na.rm=TRUE)
```

Calculate means for subgroups using the by function.

Group means, code

```
fev_means
## fev$smoke_factor: nonsmoker
## [1] 2.566143
## ------------------------------
## fev$smoke_factor: smoker
## [1] 3.276862
```

Here are the group means.

Don’t forget to round

```
round(fev_means, 1)
## fev$smoke_factor: nonsmoker
## [1] 2.6
## ------------------------------
## fev$smoke_factor: smoker
## [1] 3.3
```

More statistics, code (1/5)

```
fev_stdev <-
  by(
    fev$fev,
    fev$smoke_factor,
    sd,
    na.rm=TRUE)
```

I want to show an advanced example, for a variety of reasons. It emphasizes how in R, you take small pieces and put them together to create something complex. It also shows how to make output close to publication-ready. Finally, it explains the counter-intuitive finding the smokers have a higher average FEV than non-smokers.

You don’t need to use this extra level of effort for your homework.

First start by computing standard deviations for FEV. grouped by smoking status.

More statistics, code (2/5)

```
age_means <-
  by(
    fev$fev,
    fev$smoke_factor,
    mean,
    na.rm=TRUE)
```

Now compute the average age, grouped by smoking status.

More statistics, code (3/5)

```
age_stdev <-
  by(
    fev$fev,
    fev$smoke_factor,
    sd,
    na.rm=TRUE)
```

Compute the standard deviations for age, grouped by smoking status.

More statistics, code (4/5)

```
colon <- ": "
plus_minus <- "+/-"
fev_stats <- paste0(
  names(fev_means),
  colon,
  round(fev_means, 1),
  plus_minus,
  round(fev_stdev, 1))
```

Now combine the FEV means and standard deviations into a single string.

More statistics, code (5/5)

```
age_stats <- paste0(
  names(age_means),
  colon,
  round(age_means,1),
  plus_minus,
  round(age_stdev,1),
  sep="")
```

Now combine the age means and standard deviations the same way.

More statistics, output

```
fev_stats
## [1] "nonsmoker: 2.6+/-0.9"
## [2] "smoker: 3.3+/-0.7"
age_stats
## [1] "nonsmoker: 2.6+/-0.9"
## [2] "smoker: 3.3+/-0.7"
```

Notice that smokers are 4 years older on average than nonsmokers. Older children have bigger lungs and larger FEV values. So age is a confounding variable for the effect of smoking on FEV.

Now I need to emphasize again that you don’t have to provide this level of complexity in your homework…unless you are ambitious and want to try something a bit out of the ordinary.

Break #3

- What have you learned?
  - Review earlier material
  - Boxplots
  - Group means and standard deviations
- What's coming up next?
  - Datasets needed for your homework

Time for another break. You saw how to use boxplots to visually compare a continuous variable across different levels of a categorical variable.

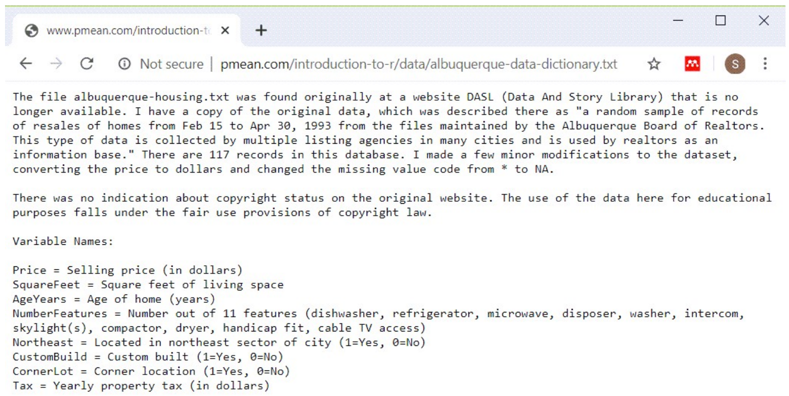
Coming up next is the calculation of group means and standard deviations.

The Albuquerque dataset

- Housing data dictionary
  - http://www.pmean.com/introduction-to-r/data/albuquerque-data-dictionary.txt
- Housing dataset
  - http://www.pmean.com/introduction-to-r/data/albuquerque-housing.txt

The first file you will need for your homework is the Albuquerque dataset. It has information on 117 housing resales in the city of Albuquerque, New Mexico back in 1993. The dataset includes information on the size and age of the house, among other things, that might be predictive of the sales price.

The Albuquerque dataset




```
The file albuquerque-housing.txt was found originally at a website DASL (Data And Story Library) that is no
longer available. I have a copy of the original data, which was described there as "a random sample of records
of resales of homes from Feb 15 to Apr 30, 1993 from the files maintained by the Albuquerque Board of Realtors.
This type of data is collected by multiple listing agencies in many cities and is used by realtors as an
information base." There are 117 records in this database. I made a few minor modifications to the dataset,
converting the price to dollars and changed the missing value code from * to NA.

There was no indication about copyright status on the original website. The use of the data here for educational
purposes falls under the fair use provisions of copyright law.

Variable Names:

Price = Selling price (in dollars)
SquareFeet = Square feet of living space
AgeYears = Age of home (years)
NumberFeatures = Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom,
skylight(s), compactor, dryer, handicap fit, cable TV access)
Northeast = Located in northeast sector of city (1=Yes, 0=No)
CustomBuild = Custom built (1=Yes, 0=No)
CornerLot = Corner location (1=Yes, 0=No)
Tax = Yearly property tax (in dollars)
```

Data dictionary for the Albuquerque dataset

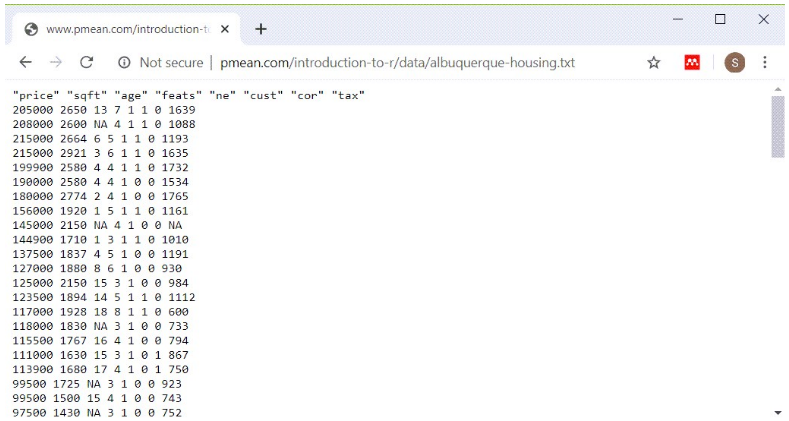
This is the data dictionary. Notice that there are several categorical variables.

Northeast is an indicator variable for whether the house was located in the Northeast part of the city, which is believed to be slightly more upscale than the other parts of the city.

CustomBuild is an indicator variable for whether the house was built using special (custom) plans or if it was built using standard plans.

CornerLot is an indicator variable for whether the house sat on a corner lot (a lot at the intersection of two streets). There are reasons to believe that houses on corner lots should be more expensive than other houses, but also reasons to believe that these houses should be less expensive.

The Albuquerque dataset




```
"price" "sqft" "age" "feats" "ne" "cust" "cor" "tax"
205000 2650 13 7 1 1 0 1639
208000 2600 NA 4 1 1 0 1088
215000 2664 6 5 1 1 0 1193
215000 2921 3 6 1 1 0 1635
199900 2580 4 4 1 1 0 1732
190000 2580 4 4 1 0 0 1534
180000 2774 2 4 1 0 0 1765
156000 1920 1 5 1 1 0 1161
145000 2150 NA 4 1 0 0 NA
144900 1710 1 3 1 1 0 1010
137500 1837 4 5 1 0 0 1191
127000 1880 8 6 1 0 0 930
125000 2150 15 3 1 0 0 984
123500 1894 14 5 1 1 0 1112
117000 1928 18 8 1 1 0 600
118000 1830 NA 3 1 0 0 733
115500 1767 16 4 1 0 0 794
111000 1630 15 3 1 0 1 867
113900 1680 17 4 1 0 1 750
99500 1725 NA 3 1 0 0 923
99500 1500 15 4 1 0 0 743
97500 1430 NA 3 1 0 0 752
```

A peek at the raw data for the Albuquerque dataset

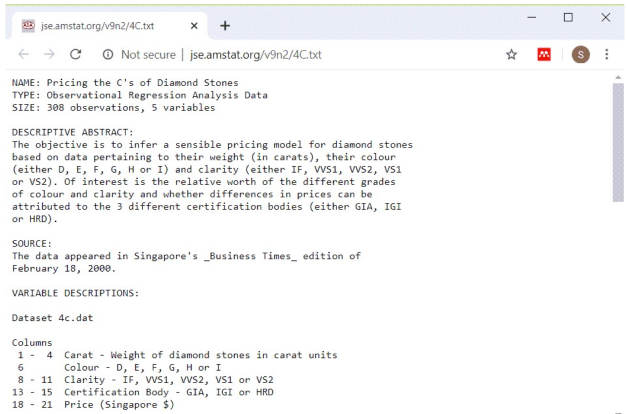
There is only a single blank between each data value. Use a blank as a delimiter.

The Diamond dataset

- Diamond data dictionary
  - http://jse.amstat.org/v9n2/4C.txt
- Diamond dataset
  - http://jse.amstat.org/v9n2/4Cdata.txt

Note that the data dictionary describes two different datasets. You will be using the first dataset, the one with a smaller number of variables.

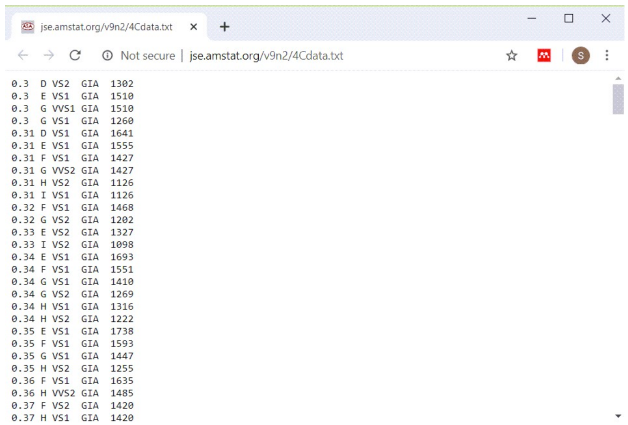
The Diamond dataset


```
NAME: Pricing the C's of Diamond Stones
TYPE: Observational Regression Analysis Data
SIZE: 308 observations, 5 variables

DESCRIPTIVE ABSTRACT:
The objective is to infer a sensible pricing model for diamond stones
based on data pertaining to their weight (in carats), their colour
(either D, E, F, G, H or I) and clarity (either IF, VVS1, VVS2, VS1
or VS2). Of interest is the relative worth of the different grades
of colour and clarity and whether differences in prices can be
attributed to the 3 different certification bodies (either GIA, IGI
or HRD).

SOURCE:
The data appeared in Singapore's _Business Times_ edition of
February 18, 2000.

VARIABLE DESCRIPTIONS:

Dataset 4c.dat

Columns
 1 -  4  Carat - Weight of diamond stones in carat units
 6       Colour - D, E, F, G, H or I
 8 - 11  Clarity - IF, VVS1, VVS2, VS1 or VS2
13 - 15  Certification Body - GIA, IGI or HRD
18 - 21  Price (Singapore $)
```

Data dictionary for the Diamond dataset

This is the data dictionary.

The Diamond dataset




```
0.3  D VS2  GIA  1302
0.3  E VS1  GIA  1510
0.3  G VVS1 GIA  1510
0.3  G VS1  GIA  1260
0.31 D VS1  GIA  1641
0.31 E VS1  GIA  1555
0.31 F VS1  GIA  1427
0.31 G VVS2 GIA  1427
0.31 H VS2  GIA  1126
0.31 I VS1  GIA  1126
0.32 F VS1  GIA  1468
0.32 G VS2  GIA  1202
0.33 E VS2  GIA  1327
0.33 I VS2  GIA  1098
0.34 E VS1  GIA  1693
0.34 F VS1  GIA  1551
0.34 G VS1  GIA  1410
0.34 G VS2  GIA  1269
0.34 H VS1  GIA  1316
0.34 H VS2  GIA  1222
0.35 E VS1  GIA  1738
0.35 F VS1  GIA  1593
0.35 G VS1  GIA  1447
0.35 H VS2  GIA  1255
0.36 F VS1  GIA  1635
0.36 H VVS2 GIA  1485
0.37 F VS2  GIA  1420
0.37 H VS1  GIA  1420
```

A peek at the raw data for the Diamond dataset

This could either be a tab delimited file or a fixed width file. If you scroll through enough of the data, you will see that most of the variables are left justified, but the price is right justified. Tab delimited files almost always have left justification for all data. So try a fixed width format.