

Practice Project 2.1 - Preparing School Data

Business Understanding

A school district wants to predict the per pupil costs of a school based on some high level summary data about the school. This way they'll have a good estimation of how well a school is managing its costs relative to what the model would predict. You've been asked to to prepare the data for modelling.

Data Understanding

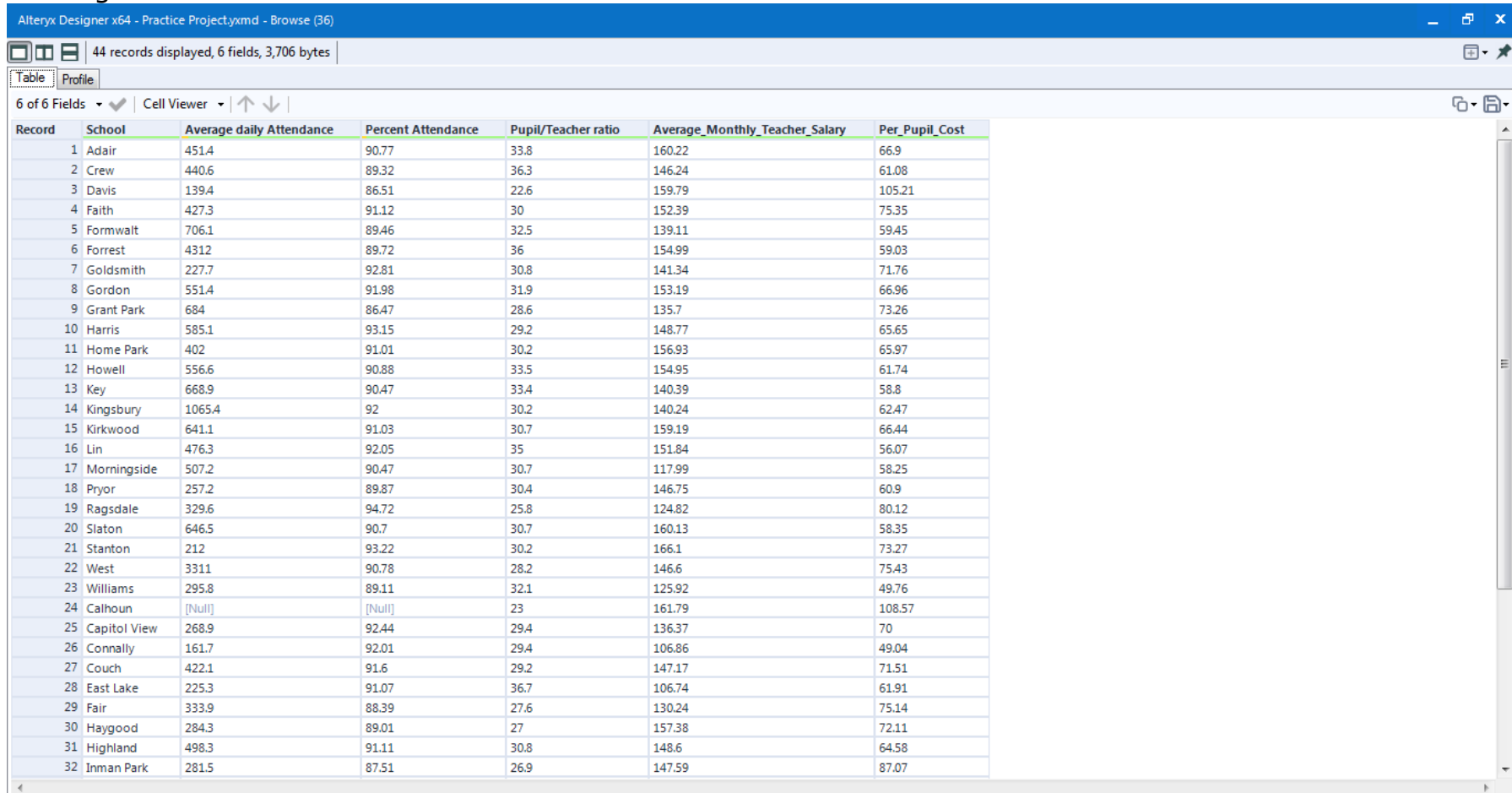
To begin, let's take a look at the data. The data is in four different csv files, which we'll have to combine in order to do the analysis. The data will have to be transformed before we can merge it with the attendance data.

- First, we'll have to transform the finance datasets
 - Then we'll merge the finance and attendance datasets for each district
 - And lastly we'll combine the data for the two districts together.
1. Transforming the data: Use Cross Tab to pivot the orientation of data in a table by moving vertical data fields onto a horizontal axis and summarizing data where specified.
 2. Merging the data: Use Join to combine 2 inputs based on common fields between the 2 tables.
 3. Combine data: Use Union to combine 2 or more datasets on column names or positions. In the output, each column contains the rows from each input.

Data Preparation

Step 1: Combine the data

First you'll need to combine the data from the various files into one sheet, with one row per school. To do this, you'll use the skills you learned in the Formatting Data and Blending Data lessons. To build the dataset, we'll have to merge each of these datasets into one.



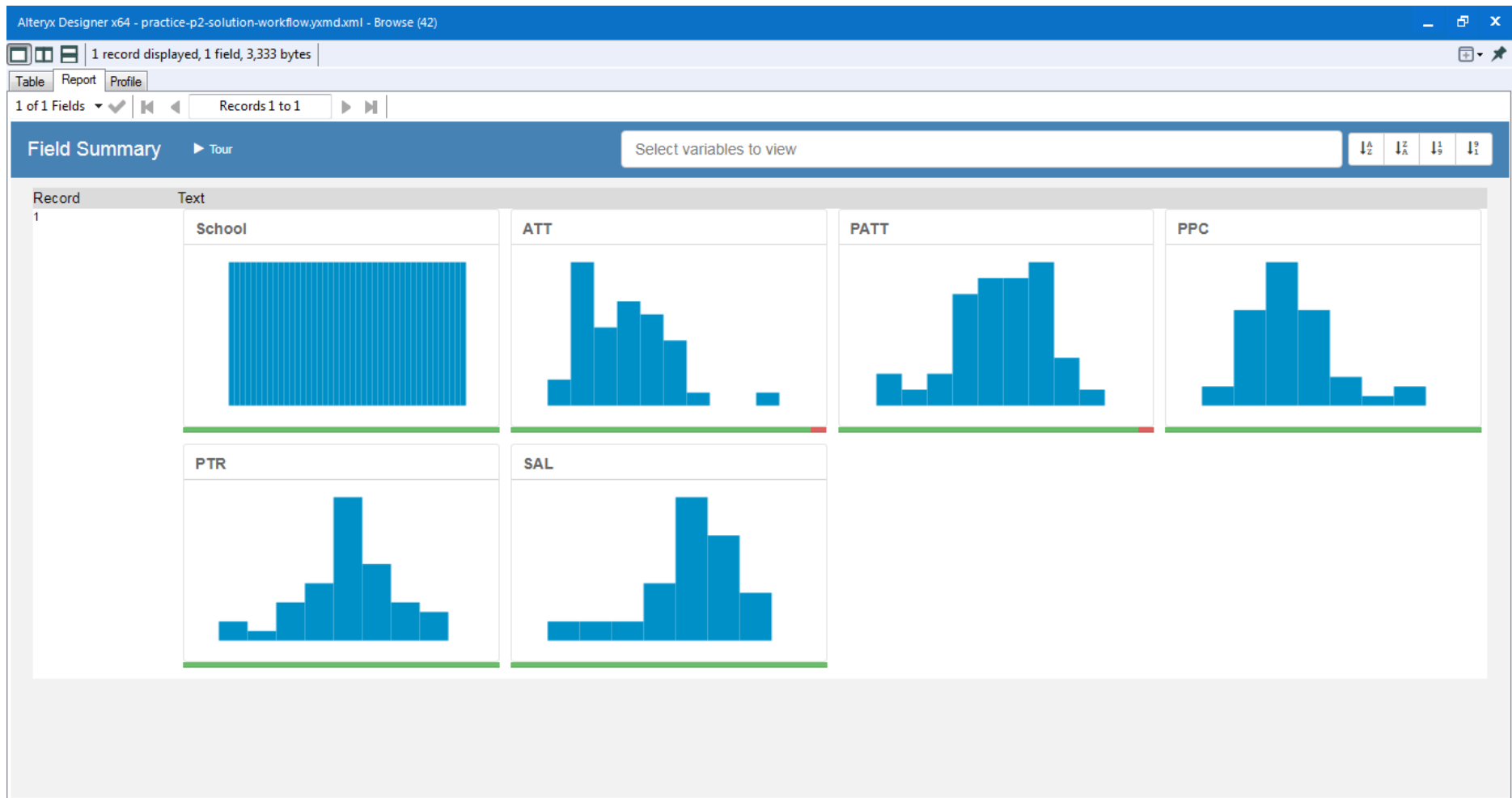
The screenshot shows the Alteryx Designer x64 interface. The title bar reads "Alteryx Designer x64 - Practice Project.yxmd - Browse (36)". Below the title bar, a status bar indicates "44 records displayed, 6 fields, 3,706 bytes". The main workspace displays a table with 32 records and 6 fields. The table is titled "Table" and has a "Profile" tab. The table columns are: Record, School, Average daily Attendance, Percent Attendance, Pupil/Teacher ratio, Average_Monthly_Teacher_Salary, and Per_Pupil_Cost. The data is sorted by Record number. The table is displayed in "Cell Viewer" mode. The table data is as follows:

Record	School	Average daily Attendance	Percent Attendance	Pupil/Teacher ratio	Average_Monthly_Teacher_Salary	Per_Pupil_Cost
1	Adair	451.4	90.77	33.8	160.22	66.9
2	Crew	440.6	89.32	36.3	146.24	61.08
3	Davis	139.4	86.51	22.6	159.79	105.21
4	Faith	427.3	91.12	30	152.39	75.35
5	Formwalt	706.1	89.46	32.5	139.11	59.45
6	Forrest	4312	89.72	36	154.99	59.03
7	Goldsmith	227.7	92.81	30.8	141.34	71.76
8	Gordon	551.4	91.98	31.9	153.19	66.96
9	Grant Park	684	86.47	28.6	135.7	73.26
10	Harris	585.1	93.15	29.2	148.77	65.65
11	Home Park	402	91.01	30.2	156.93	65.97
12	Howell	556.6	90.88	33.5	154.95	61.74
13	Key	668.9	90.47	33.4	140.39	58.8
14	Kingsbury	1065.4	92	30.2	140.24	62.47
15	Kirkwood	641.1	91.03	30.7	159.19	66.44
16	Lin	476.3	92.05	35	151.84	56.07
17	Morningside	507.2	90.47	30.7	117.99	58.25
18	Pryor	257.2	89.87	30.4	146.75	60.9
19	Ragsdale	329.6	94.72	25.8	124.82	80.12
20	Slaton	646.5	90.7	30.7	160.13	58.35
21	Stanton	212	93.22	30.2	166.1	73.27
22	West	3311	90.78	28.2	146.6	75.43
23	Williams	295.8	89.11	32.1	125.92	49.76
24	Calhoun	[Null]	[Null]	23	161.79	108.57
25	Capitol View	268.9	92.44	29.4	136.37	70
26	Connally	161.7	92.01	29.4	106.86	49.04
27	Couch	422.1	91.6	29.2	147.17	71.51
28	East Lake	225.3	91.07	36.7	106.74	61.91
29	Fair	333.9	88.39	27.6	130.24	75.14
30	Haygood	284.3	89.01	27	157.38	72.11
31	Highland	498.3	91.11	30.8	148.6	64.58
32	Inman Park	281.5	87.51	26.9	147.59	87.07

Step 2: Clean the Data

Next you'll clean the data, which includes addressing duplicate data, missing data, and any other data issues. To do this, you'll use the skills you learned in the Data Issues lesson.

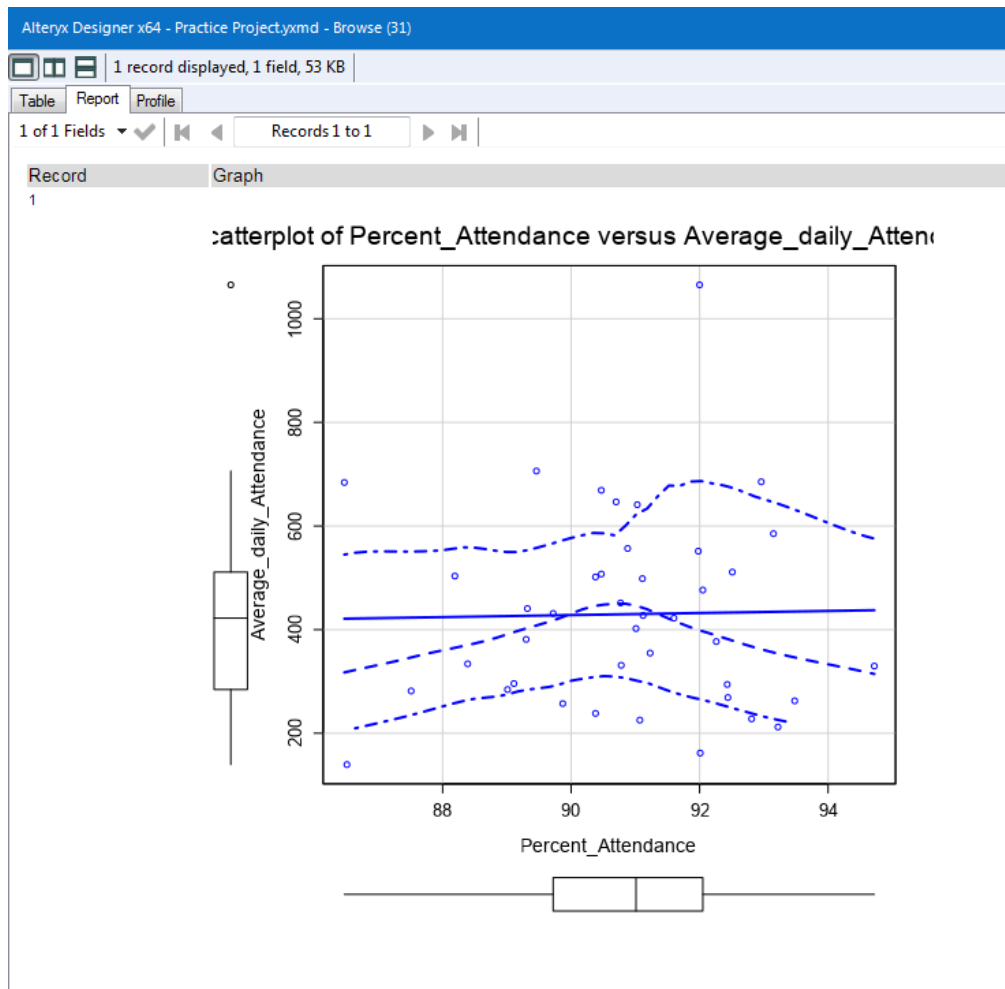
Visualizing the data is a good way of doing this. The field summary report gives a few reports that are helpful. Looking at histograms of each variable is a good way to do this.



Step 3: Identify and Deal with Outliers

Lastly, you'll look for outliers and determine the best way to address them. To do this, you'll use the skills you learned in the Data Issues lesson.

NOTE: In this example, because of the small size of the dataset, we could look at each outlier and make decisions. For larger data sets, you'd likely have to make more systematic decisions, such as removing all outliers, or removing the top 1 or 2 percent of observations for each variable.



Modeling

The Linear Regression Tool creates a simple model to estimate values, or evaluate relationships between variables based on a linear relationship.

Alteryx Designer x64 - practice-p2-solution-workflow.yxmd.xml - Browse (79)

12 records displayed, 2 fields, 101 KB

Table Report Profile

1 of 1 Fields ▾

Records 1 to 10

2	Basic Summary				
3	Call: lm(formula = PPC ~ ATT + PTR + SAL, data = the.data)				
4	Residuals:				
5	Min	1Q	Median	3Q	Max
	-16.812	-3.186	0.156	3.732	15.684
6	Coefficients:				
7		Estimate	Std. Error	t value	Pr(> t)
	(Intercept)	107.14271	16.056845	6.673	7.76e-08 ***
	ATT	-0.01593	0.006278	-2.538	0.01549 *
	PTR	-2.22133	0.378135	-5.874	9.28e-07 ***
	SAL	0.24885	0.076601	3.249	0.00247 **
	Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
8	Residual standard error: 6.9211 on 37 degrees of freedom Multiple R-squared: 0.6306, Adjusted R-Squared: 0.6007 F-statistic: 21.06 on 3 and 37 degrees of freedom (DF), p-value 3.969e-08				
9	Type II ANOVA Analysis				
10	Response: PPC				
		Sum Sq	DF	F value	Pr(>F)
	ATT	308.57	1	6.44	0.01549 *
	PTR	1653.03	1	34.51	9.28e-07 ***
	SAL	505.56	1	10.55	0.00247 **
	Residuals	1772.35	37		
	Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				