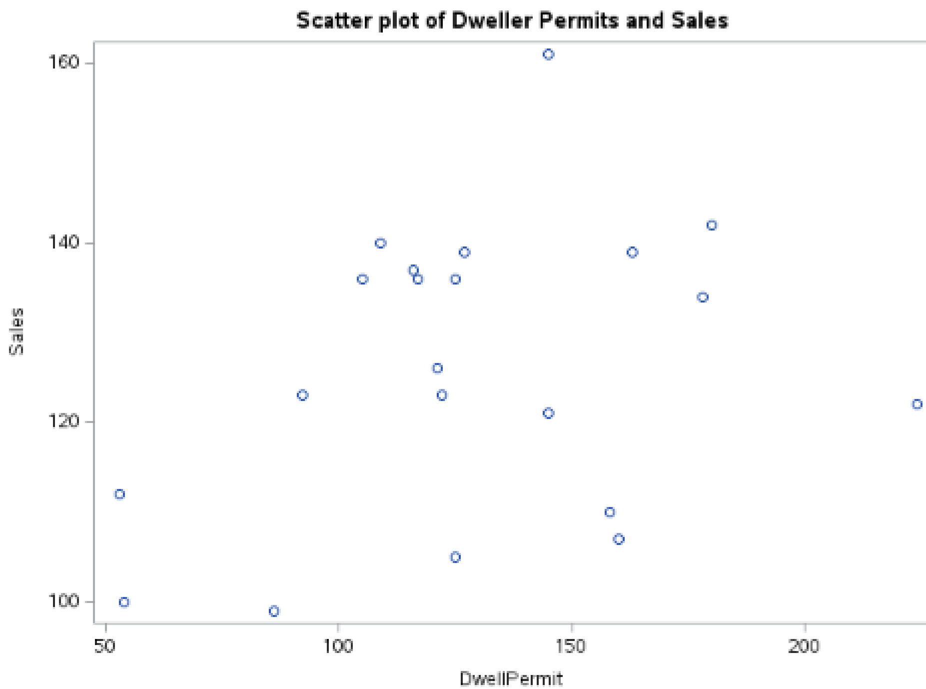
DATA 3441: Lab 2 Submission

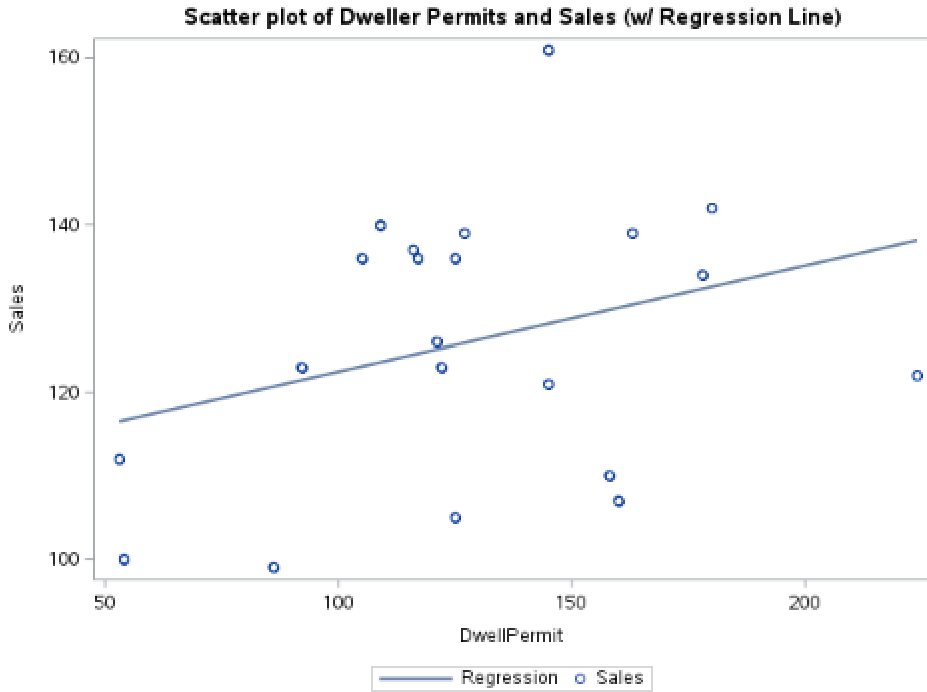
1. (2.144, Dwelling permits and sales for 23 countries.)

(a)



There may be a positive linear relationship between the amount of dwelling permits affecting sales. I think there are two outliers at y={160,125}. It may be hard to see but the majority of sales are between 100 and 150 permits.

(b)



The same graph with the least-squares regression line added.

(c)

$$b_1 = r(\sigma_y/\sigma_x)$$

The term b_1 , our slope, represents an increase of ~109.8204 dwelling permits for every sale made.

(d)

$$b_0 = \bar{y} - b_1(X)$$

b_0, our intercept, predicts that there were 0.1263 dwelling permits at the start of data collection, when sales are at zero. In the context of the exercise, that would mean that at the start of year 2000, there were nearly zero dwelling permits on average.

(e)

$$\hat{y} = b_0 + b_1(x)$$

At the index of 224, our predicted sales value is 138.1223 sales when a country has 224 permits.

(f)

$$\text{Residual} = (y-\hat{y}) = \text{Error}$$

The residual, or error for the sales is about 85.8777 sales.

(g)

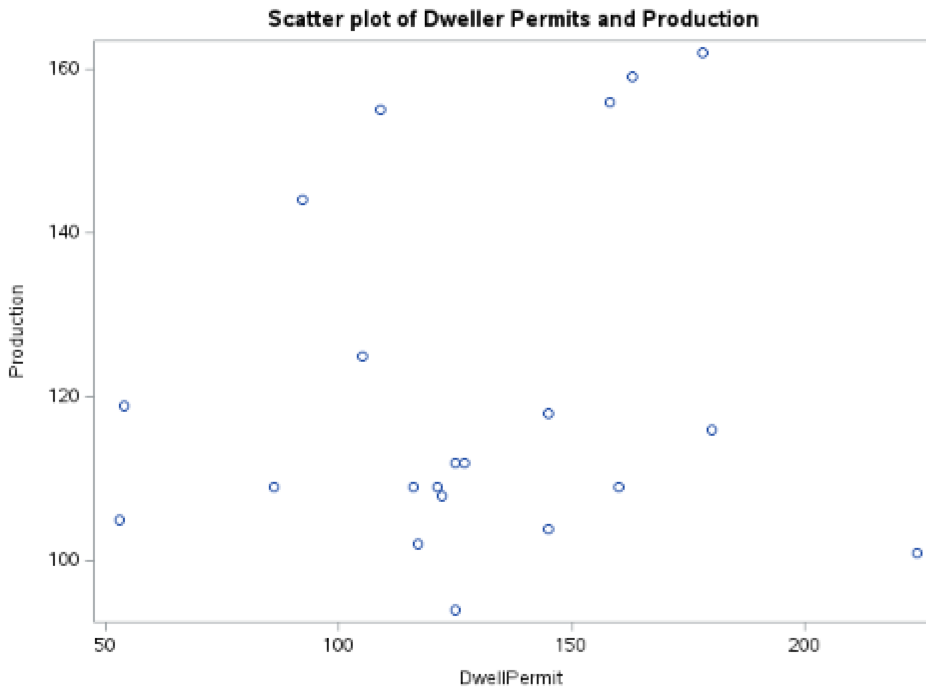
$$r = (1/n-1)\sum[z_{xi} * z_{yi}]$$

$$r^2 = r*r$$

The coefficient of determination reveals that 10.26 percent (r^2 = 0.1026) of sales variation explains dwelling permits.

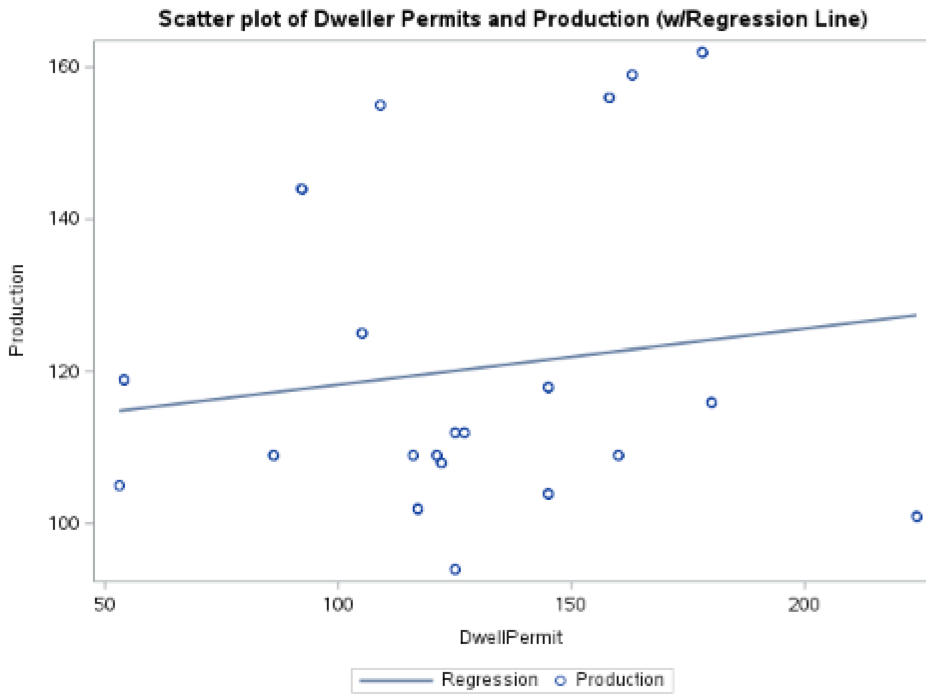
2. (2.145, Dwelling permits and production.)

(a)



Looking at the scatter plot without a line, it is unclear what the relationship is. If I were to make an inference, there is a slight positive relationship between dwelling permits and production. There may be one outlier at x=~250.

(b)



The same plot with the least-squares regression line added.

(c)

There are nearly zero changes in dwelling permits ($b_1 = 0.0731$) for every production.

(d)

When there are zero permits, the main economic indicator (MEI) for the 23 countries listed was 110.9581 permits when production was at zero.

(e)

The predicted value of production for a country with index 224 was 127.3325 productions.

(f)

The residual for Canada with an index of 224 was 96.6675 permits.

(g)

The coefficient of determination reveals that 1.98 percent (r^2 = 0.0198) of production variation explains dwelling permits. Compared to the value of 10.26 percent (r^2 = 0.1026) for sales variation, we can hypothesize that sales has a stronger effect on dwelling permits than production.

3. (2.153, Fields of study for college students.)

(a)

Field of Study	Canada	France	Germany	Italy	Japan	UK	US	Total
Social Sciences	64	153	66	125	250	152	878	1688
STEM	35	111	66	80	136	128	355	911
Art	27	74	33	42	123	105	397	801
Education	20	45	18	16	39	14	167	319
Other	30	289	35	58	97	76	272	857
Total	176	672	218	321	645	475	2069	4576

(b)

Country	Percent $n=4576$
Canada	$176/n = 0.0385 \Rightarrow 3.85\%$
France	$672/n = 0.1469 \Rightarrow 14.69\%$
Germany	$218/n = 0.0476 \Rightarrow 4.76\%$
Italy	$321/n = 0.0682 \Rightarrow 6.82\%$
Japan	$645/n = 0.1410 \Rightarrow 14.10\%$
UK	$475/n = 0.1038 \Rightarrow 10.38\%$
US	$2069/n = 0.4521 \Rightarrow 45.21\%$

Marginal Distribution of Countries

(c)

Field of Study	Percent $n = 4576$
Social Sciences	$1688/n = 0.3689 \Rightarrow 36.89\%$
STEM	$911/n = 0.1991 \Rightarrow 19.91\%$
Art	$801/n = 0.1750 \Rightarrow 17.50\%$
Education	$319/n = 0.0697 \Rightarrow 6.97\%$
Other	$857/n = 0.1873 \Rightarrow 18.73\%$

Marginal Distribution of Field of Study

4. (2.154. Fields of study by country for college students.)

(a and b)

Distributions of graduates in each country in Fields of Study (Conditional)

Field of Study	Canada	France	Germany	Italy	Japan	U.K	U.S	Total
Social Sciences	64/n = 36.36	153/n = 22.77	66/n = 30.28	125/n = 38.94	250/n = 38.76	152/n = 32.00	878/n = 42.44	1688
STEM	35/n = 19.87	111/n = 16.52	66/n = 30.28	80/n = 24.92	136/n = 21.08	128/n = 26.95	355/n = 17.16	911
Art	27/n = 15.34	74/n = 11.01	33/n = 15.14	42/n = 13.08	123/n = 19.07	105/n = 22.11	397/n = 19.19	801
Education	20/n = 11.36	45/n = 6.70	18/n = 8.26	16/n = 4.98	39/n = 6.05	14/n = 2.95	167/n = 8.07	319
Other	30/n = 17.05	289/n = 43.00	35/n = 16.06	58/n = 18.07	97/n = 15.04	76/n = 16.00	272/n = 13.15	857
Total(n)	176	672	218	321	645	475	2069	4576

5. (2.156, Salaries and raises.)

(a)

It seems that the longer the employee is employed at xyz company, her salary increases by $2,000 each year.

It seems that the longer the employee stays at the company, the salary increases about $2,000 each year.

(b)

The coefficient of determination in this case represents the proportion of salary that is explained by year. The percent is 98.32 ($r^2 = 0.9832$) of the variance is explained by salary. The high percentage indicates a strong linear relationship between salary and years worked, which means our prediction is very likely to be true.

6. (2.158, Try logs.)

(a)

For figure 2.35, there is a curvilinear correlation, which indicates a relationship is there, just that it is not linear. The LSRL line is at zero, so there is no slope, which means there is no relationship that is linear. This would suggest that a correlation cannot be made and that a transformation would have to happen.

(b)

Figure 2.37 has its points scattered, which is better for our predictions. The log transformation reduced the values into data that is more interpretable and standardized.