Exploratory Data Analysis - Part Two

Summary

- Frequency Distribution.
- Correlation Analysis.
- Scatter Plot.
- Regression Analysis.
- Analysis of Variance (ANOVA).

Frequency Distribution

- It consists of grouping the data into categories that show the number of cases or observations in each mutually exclusive category.

500	3000	2500	680	550
900	1400	750	850	2500
900	650	1320	700	1300
1500	2500	240	1900	750
1300	900	800	2100	2050
600	1350	1100	750	1400
1400	1900	950	800	900
2000	700	630	1000	600

Frequency Distribution

- Step 1: Establish categorical groups called classes.
- Step 2: Distribute the data in the corresponding class.
- Step 3: Count the amount of data in each class.

Frequency Distribution

- Step 1: Establish categorical groups called classes.
  - Define the class interval.

$$\text{class interval} = \frac{\text{Maximum value - Minimum value}}{\text{number of classes}}$$

Frequency Distribution

- Step 1: Establish categorical groups called classes.
  - Define the class interval.

$$\text{class interval} = \frac{\text{Maximum value - Minimum value}}{\text{number of classes}}$$

  - Rules of thumb for determining the number of classes.

1) Not less than 5 and not more than 15.
2) $2^k >= n$ ($k = 0,1,2\ldots$) (n = number of observations).

k	2^k	n
0	1	40
1	2	40
2	4	40
3	8	40
4	16	40
5	32	40
6	64	40

$$\text{class interval} = \frac{3{,}000 - 240}{6} = \mathbf{460}$$

Salaries	
0	499
500	999
1,000	1,499
1,500	1,999
2,000	2,499
2,500	3,000

Frequency Distribution

- Step 2: Distribute the data in the corresponding class.

Salaries	
0	499
500	999
1,000	1,499
1,500	1,999
2,000	2,499
2,500	3,000

240
500
550
600
600
630
650
680
700
700
750
750
750
800
800
850
900
900
900
900
950
1000

Frequency Distribution

- Step 3: Count the amount of data in each class.

Salaries	
0	499
500	999
1,000	1,499
1,500	1,999
2,000	2,499
2,500	3,000

240	
500	1
550	2
600	3
600	4
630	5
650	6
680	7
700	8
700	9
750	10
750	11
750	12
800	13
800	14
850	15
900	16
900	17
900	18
900	19
950	20
1000	

SALARIES		
From	To	Frequency
0	499	1
500	999	20
1,000	1,499	9
1,500	1,999	3
2,000	2,499	3
2,500	3,000	4

Frequency Distribution in Python

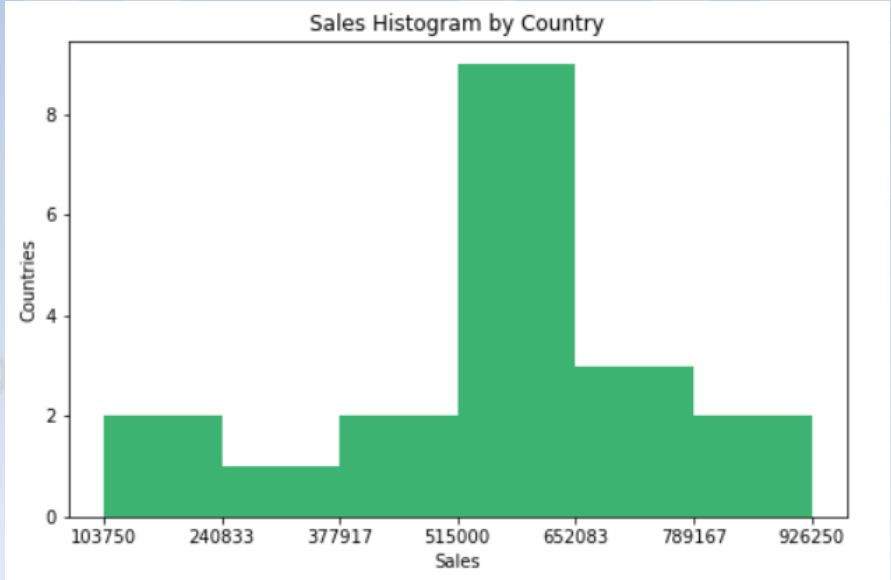
- Histogram.
  - plot() method.

```
Country
Canada          150000
Japan           651750
Mexico          563750
Spain           706250
Argentina       640375
Canada          519375
EEUU            870375
Chile           926250
EEUU            676250
EEUU            103750
Name: Sales, dtype:
```

```python
count, bin_edges = np.histogram(df_operations['Sales'], num_bin)

df_operations['Sales'].plot(kind = 'hist',
                            figsize = (8,5),
                            bins = bin_edges,
                            xticks = bin_edges,
                            color ='mediumseagreen'
                           )

plt.title('Sales Histogram by Country')
plt.ylabel('Countries')
plt.xlabel('Sales')
plt.show()
```



Correlation Analysis

- Measures the strength of correlation between two variables.
  - Correlation coefficient
    - Close to 1 : High positive correlation
    - Close to -1 : High negative correlation
    - Close to 0 : No correlation
  - p value
    - $p < 0.001$: High certainty in the result
    - $p < 0.05$: Moderate certainty in the result
    - $p < 0.1$: Low certainty in the result
    - $p > 0.1$: Lack of certainty in the result

    https://en.wikipedia.org/wiki/P-value

Correlation Analysis

❑ Pearson's Correlation Coefficient.

$$r = \frac{\sum X\gamma - n(\bar{x})\,(\bar{y})}{\sqrt{\left(\sum X^2 - n\bar{X}^2\right)\left(\sum y^2 - n\bar{y}^2\right)}}$$

$$r = \frac{10{,}143{,}130{,}842 - 19(563{,}252)(819)}{\sqrt{(6{,}814{,}828{,}787{,}298 - (19)(563{,}252)^2)(16{,}227{,}383 - (19)(819)^2)}}$$

$$r = 0.83195$$

	X	Y
mean	563,252	819

SALES	REFUNDS				
X	Y	XY	X^2	Y^2
150,000	240	36,000,000	22,500,000,000	57,600
651,750	1,043	679,775,250	424,778,062,500	1,087,849
563,750	902	508,502,500	317,814,062,500	813,604
706,250	1,130	798,062,500	498,789,062,500	1,276,900
640,375	1,024	655,744,000	410,080,140,625	1,048,576
519,375	0	0	269,750,390,625	0
870,375	1,392	1,211,562,000	757,552,640,625	1,937,664
926,250	1,482	1,372,702,500	857,939,062,500	2,196,324
676,250	1,082	731,702,500	457,314,062,500	1,170,724
103,750	166	17,222,500	10,764,062,500	27,556
567,925	910	516,811,750	322,538,805,625	828,100
650,041	1,041	676,692,681	422,553,301,681	1,083,681
565,000	904	510,760,000	319,225,000,000	817,216
440,000	704	309,760,000	193,600,000,000	495,616
301,262	480	144,605,760	90,758,792,644	230,400
700,152	1,120	784,170,240	490,212,823,104	1,254,400
452,750	0	0	204,982,562,500	0
565,287	903	510,454,161	319,549,392,369	815,409
651,250	1,042	678,602,500	424,126,562,500	1,085,764
10,701,792	15,565	10,143,130,842	6,814,828,787,298	16,227,383

Correlation Analysis in Python

- Correlation Coefficient with p value.

	Customer	Customer Type	Payment Type	Purchases	Sales	Refunds	Country	Continent
0	10000	Person	Cash	120000	150000	240	Canada	America
1	10001	Company	Cash	521400	651750	1043	Japan	Asia
2	10002	Company	Credit Card	451000	563750	902	Mexico	America
3	10003	Company	Transfer	565000	706250	1130	Spain	Europe
4	10004	Person	Transfer	512300	640375	1024	Argentina	America

```
from scipy import stats
```

```
pearson_coef, p_value = stats.pearsonr(df_operations['Sales'], df_operations['Refunds'])
print("Pearson's correlation coefficient: ", pearson_coef, "p value: ", p_value)
```

```
Pearson's correlation coefficient:  0.8319496410228725 p value:  1.0035802366568795e-05
High certainty
```

Correlation Analysis in Python

- Correlation Matrix.

```
df_operations[["Sales", "Refunds"]].corr()
```

	Sales	Refunds
Sales	1.00000	0.83195
Refunds	0.83195	1.00000

Correlation Analysis in Python

- Correlation Matrix.

```
df_operations[["Sales", "Refunds"]].corr()
```

	Sales	Refunds
Sales	1.00000	0.83195
Refunds	0.83195	1.00000

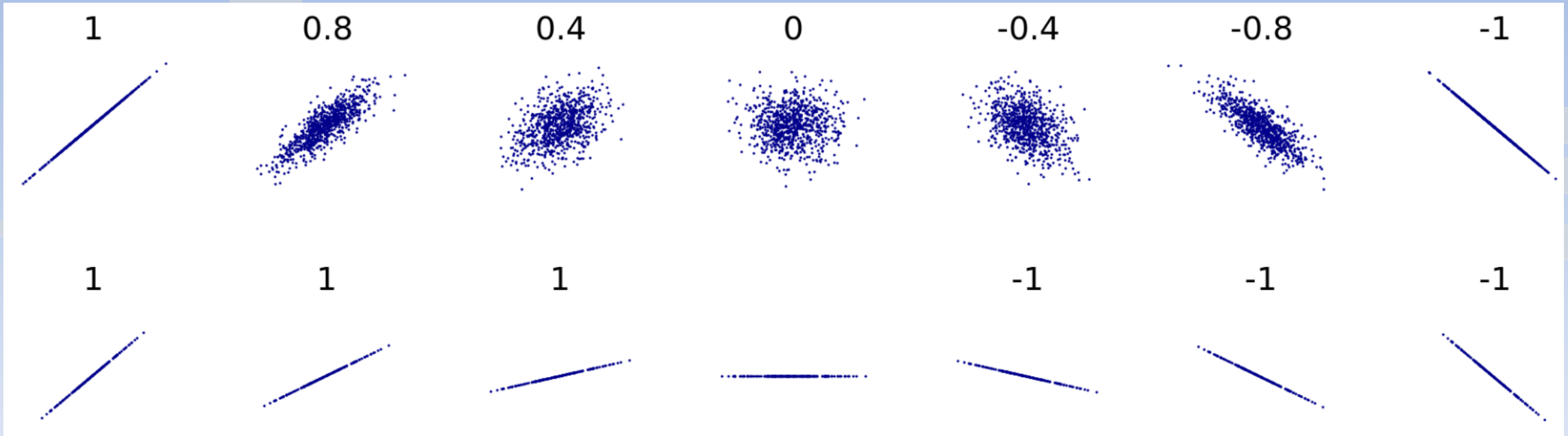
- Determination Coefficient :
  - $r^2 = (0.83195)^2 = 0.6921 = 69.21\%$

Scatter Plot

- A scatter plot is a graphical illustration that is used in regression analysis.

- It consists of a dispersion of points where each point represents a value of the independent variable (measured on the horizontal axis), and a value associated with the dependent variable (measured on the vertical axis).

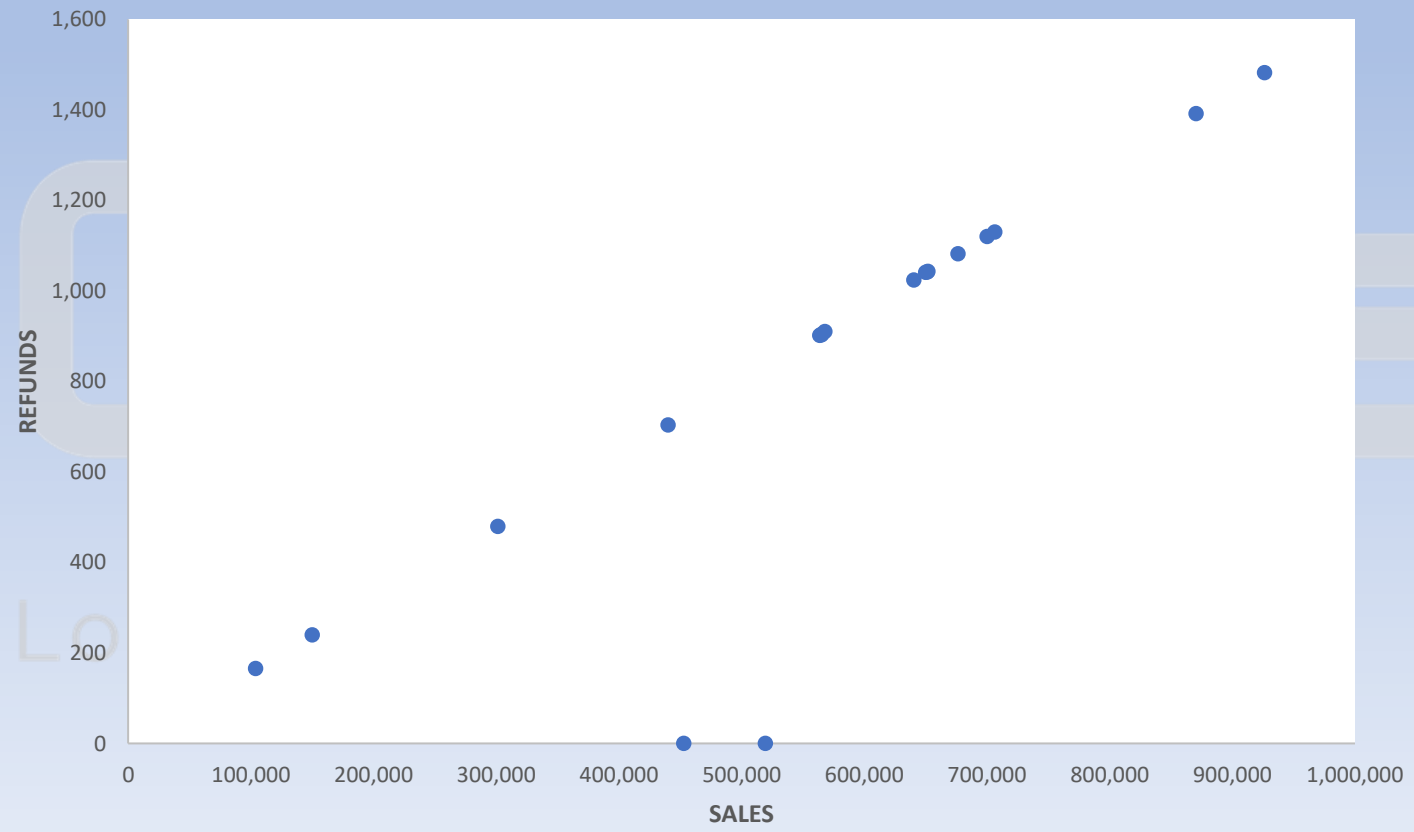
Scatter Plot



Source: *https://en.wikipedia.org/wiki/Correlation_and_dependence*

Scatter Plot

SALES	REFUNDS
X	Y
150,000	240
651,750	1,043
563,750	902
706,250	1,130
640,375	1,024
519,375	0
870,375	1,392
926,250	1,482
676,250	1,082
103,750	166
567,925	910
650,041	1,041
565,000	904
440,000	704
301,262	480
700,152	1,120
452,750	0
565,287	903
651,250	1,042
10,701,792	15,565



Regression Analysis

❑ Least squares method

$$\hat{\gamma} = a + bX$$

$$b = \frac{\sum X\gamma - n(\bar{x})(\bar{y})}{\sum X^2 - n\bar{X}^2}$$

$$a = \bar{Y} - b\bar{X}$$

SALES	REFUNDS		
X	Y	XY	X^2
150,000	240	36,000,000	22,500,000,000
651,750	1,043	679,775,250	424,778,062,500
563,750	902	508,502,500	317,814,062,500
706,250	1,130	798,062,500	498,789,062,500
640,375	1,024	655,744,000	410,080,140,625
519,375	0	0	269,750,390,625
870,375	1,392	1,211,562,000	757,552,640,625
926,250	1,482	1,372,702,500	857,939,062,500
676,250	1,082	731,702,500	457,314,062,500
103,750	166	17,222,500	10,764,062,500
567,925	910	516,811,750	322,538,805,625
650,041	1,041	676,692,681	422,553,301,681
565,000	904	510,760,000	319,225,000,000
440,000	704	309,760,000	193,600,000,000
301,262	480	144,605,760	90,758,792,644
700,152	1,120	784,170,240	490,212,823,104
452,750	0	0	204,982,562,500
565,287	903	510,454,161	319,549,392,369
651,250	1,042	678,602,500	424,126,562,500
10,701,792	15,565	10,143,130,842	6,814,828,787,298

Replacing:

$$b = \frac{10{,}143{,}130{,}842 - (19)(563{,}252)(819)}{6{,}814{,}828{,}787{,}298 - (19)(563{,}252)^2} = \mathbf{0.00175}$$

$$\mathbf{a} = 819 - 0.000175\,(563{,}252) = \mathbf{-165.64}$$

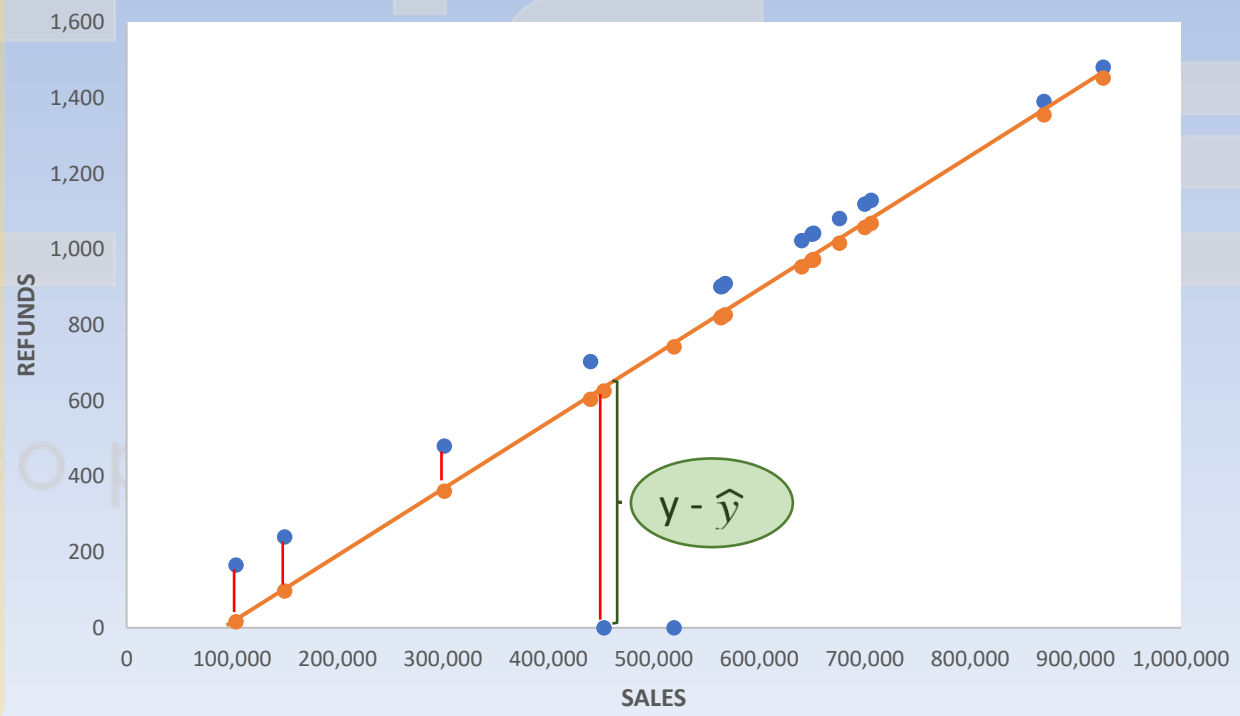
$$\hat{y} = -165.64 + 0.00175\,X$$

	X	Y
mean	563,252	819

Regression Analysis

- Least squares method.

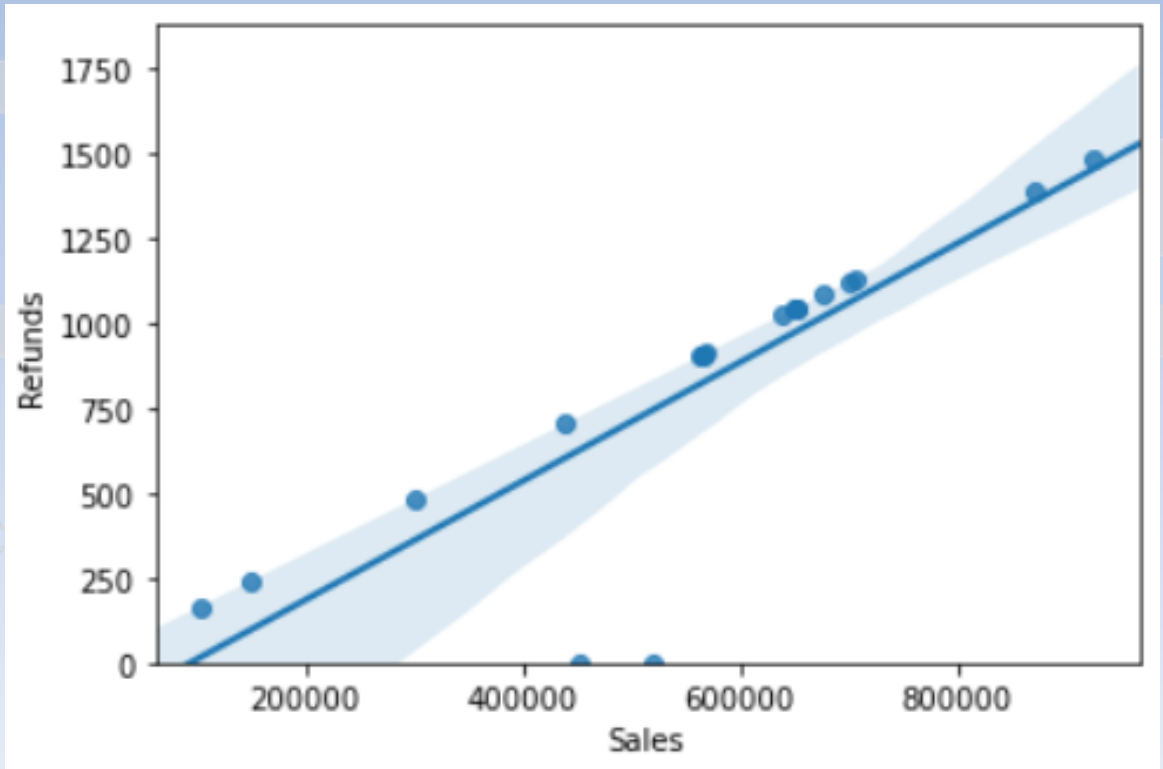
SALES	REFUNDS			
X	Y	XY	X^2	$\hat{y}$
150,000	240	36,000,000	22,500,000,000	97
651,750	1,043	679,775,250	424,778,062,500	974
563,750	902	508,502,500	317,814,062,500	820
706,250	1,130	798,062,500	498,789,062,500	1,069
640,375	1,024	655,744,000	410,080,140,625	954
519,375	0	0	269,750,390,625	742
870,375	1,392	1,211,562,000	757,552,640,625	1,356
926,250	1,482	1,372,702,500	857,939,062,500	1,454
676,250	1,082	731,702,500	457,314,062,500	1,017
103,750	166	17,222,500	10,764,062,500	16
567,925	910	516,811,750	322,538,805,625	827
650,041	1,041	676,692,681	422,553,301,681	971
565,000	904	510,760,000	319,225,000,000	822
440,000	704	309,760,000	193,600,000,000	604
301,262	480	144,605,760	90,758,792,644	361
700,152	1,120	784,170,240	490,212,823,104	1,059
452,750	0	0	204,982,562,500	626
565,287	903	510,454,161	319,549,392,369	823
651,250	1,042	678,602,500	424,126,562,500	973
10,701,792	15,565	10,143,130,842	6,814,828,787,298	



Regression Plot in Python

- regplot()

```
sns.regplot(x="Sales", y="Refunds", data=df_operations)
```



Analysis of Variance

- Statistical test used to find differences between groups of a categorical variable.
- The F value: Variation between the group means divided by the variation within the groups.
- The critical value for the distribution F.

Analysis of Variance

CustomerID	Type	Sales
10000	Person	150,000
10001	Company	651,750
10002	Company	563,750
10003	Company	706,250
10004	Person	640,375
10005	Person	519,375
10006	Company	870,375
10007	Person	926,250
10008	Company	676,250
10009	Company	103,750
10010	Company	567,925
10011	Person	650,041
10012	Person	565,000
10013	Person	440,000
10014	Company	301,262
10015	Company	700,152
10016	Person	452,750
10017	Person	565,287
10018	Company	651,250

ANOVA

Variation Source	Sum of Squares	Degree Freedom	Mean Square	F
Treatments	SST	k - 1	SST / (k - 1) = MST	MST / MSE
Error	SSE	n - k	SSE / (n - k) = MSE	
Total	SS Total	n - 1		

Where :

SS Total : Sum Squares Total
SST : Sum Squares of Treatments
SSE : Sum Squares of Error
k : Treatments number
n : Observations number
MST : Mean Square of Treatments
MSE : Mean Square of Error

$$\text{SS Total} = \sum x^2 - \frac{\left(\sum X\right)^2}{n}$$

$$\text{SST} = \sum \left(\frac{T_c^2}{n_c}\right) - \frac{\left(\sum X\right)^2}{n}$$

$$\text{SSE} = \text{SS Total} - \text{SST}$$

Analysis of Variance

	Treatment 1 Company		Treatment 2 Person		Total
	X	X^2	X	X^2	
1	652	424,778	150	22,500	
2	564	317,814	640	410,080	
3	706	498,789	519	269,750	
4	870	757,553	926	857,939	
5	676	457,314	650	422,553	
6	104	10,764	565	319,225	
7	568	322,539	440	193,600	
8	301	90,759	453	204,983	
9	700	490,213	565	319,549	
10	651	424,127			
T_c	5,793		4,909		10,702
n_c	10		9		19
X^2		3,794,649		3,020,180	6,814,829

SS Total = $6{,}814{,}829 - \frac{10{,}702^2}{19} = 787{,}020.79$

SST = $\frac{5{,}793^2}{10} + \frac{4{,}909^2}{9} - \frac{10{,}702^2}{19} = 5{,}417.42$

SSE = $787{,}020.79 - 5{,}417.42 = 781{,}603.37$

Degree Freedom: $k - 1 = 2 - 1 = 1$ (numerator)

$n - k = 19 - 2 = 17$ (denominator)

Analysis of Variance

ANOVA

Variation Source	Sum of Squares	Degree Freedom	Mean Square	F
Treatments	5,417.42	1	5,417.42	**0.11783**
Error	781,603.37	17	45,976.67	
Total	787,020.79	18		

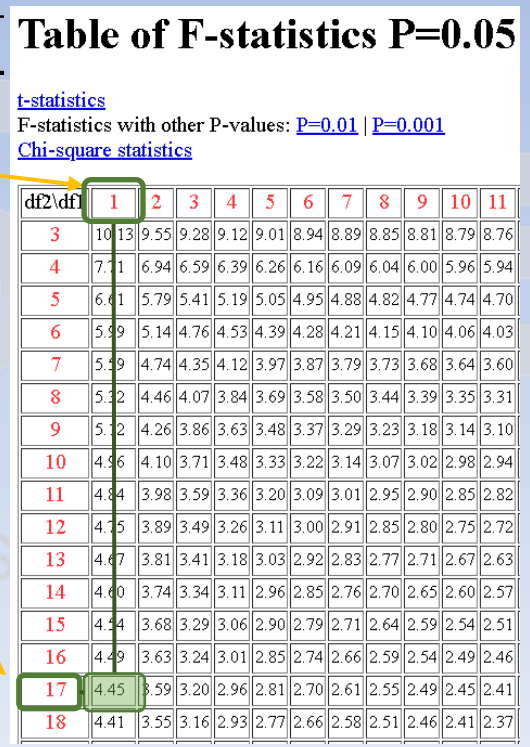


Table of F-statistics P=0.05

t-statistics
F-statistics with other P-values: P=0.01 | P=0.001
Chi-square statistics

df2\df1	1	2	3	4	5	6	7	8	9	10	11
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37

F is less than the critical value. That is, 0.11783 < 4.451.

We conclude that the average sales between the groups: company and person are equal.

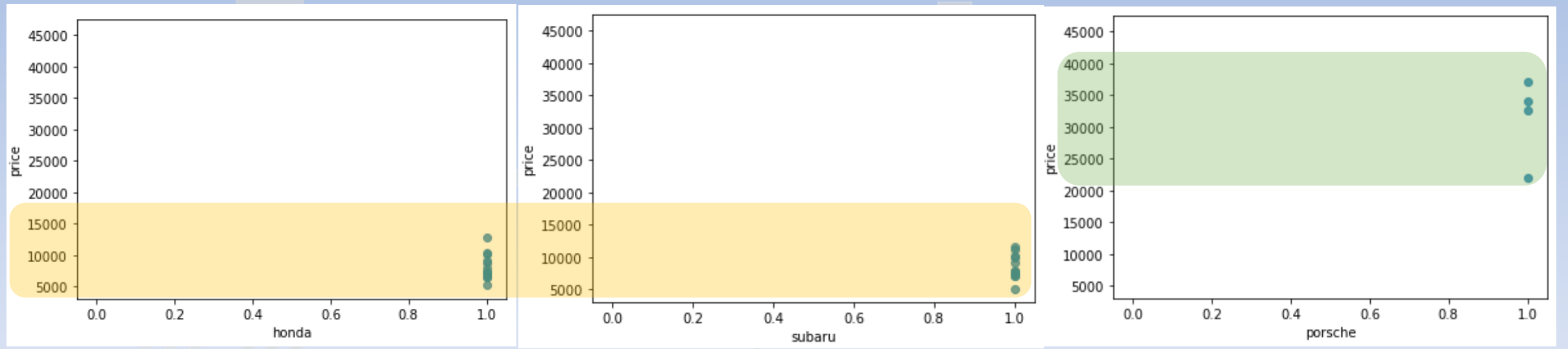
https://web.ma.utexas.edu/users/davis/375/popecol/tables/f005.html

Analysis of Variance with Python

- The F value: Variation between the group means divided by the variation within the groups.
- p value that refers to the confidence level of the test.

Analysis of Variance with Python

Scatter plot of three vehicle brands and their price



Analysis of Variance with Python

- f_oneway()

```
grouped_anova = df_anova.groupby("make")
grouped_anova.head()
```

```
anova_results_1 = stats.f_oneway(grouped_anova.get_group("honda")["price"],
                                 grouped_anova.get_group("isuzu")["price"])
anova_results_1
```

```
F_onewayResult(statistic=0.200816411197416, pvalue=0.6614400721692544)
```

```
anova_results_2 = stats.f_oneway(grouped_anova.get_group("isuzu")["price"],
                                 grouped_anova.get_group("porsche")["price"])
anova_results_2
```

```
F_onewayResult(statistic=19.685519109339, pvalue=0.011361720842259 2)
```

The F statistic is less than the critical value when you compare Honda's prices to Isuzu, but it is much higher if you compare Isuzu to Porsche.

	make	price
0	alfa-romero	13495.0
1	alfa-romero	16500.0
2	alfa-romero	16500.0
3	audi	13950.0
4	audi	17450.0
...	...	...
190	volvo	12940.0
191	volvo	13415.0
192	volvo	15985.0
193	volvo	16515.0
194	volvo	18420.0