Data Mining

Introduction

- Data Mining is about finding useful information from the data

Introduction

Data Mining is about finding useful information from the data to answer

- What happened?
- Why did it happen?
- What will happen?
- What can we do to make things happen?

How to find useful information?

by finding

- relationships between variables
- trends
- splitting the data by categorical variables
- transforming variables
- encoding variables –feature engineering–
- creating new variables

Relationship between variables

- Correlation between y and x_1 is r = 0.7

Relationship between variables

- Correlation between y and x_1 is r = 0.7
- Regression line is $\hat{y} = 0.934 + 2.114\,x_1$
- On average

  y increases by 2.114 when *x* increases by 1

Relationship between variables

- Correlation between y and x_1 is r = 0.7

  On avg, y increases when x increases

- Regression line is $\hat{y} = 0.934 + 2.114\,x_1$

  On avg, *y* increases by 2.114 when *x* increases by 1

- Add variable x_2 and now the regression line is

$$\hat{y} = 0.934 - 0.25\,x_1 + 1.76\,x_2$$

  What is the relation between y and x1?

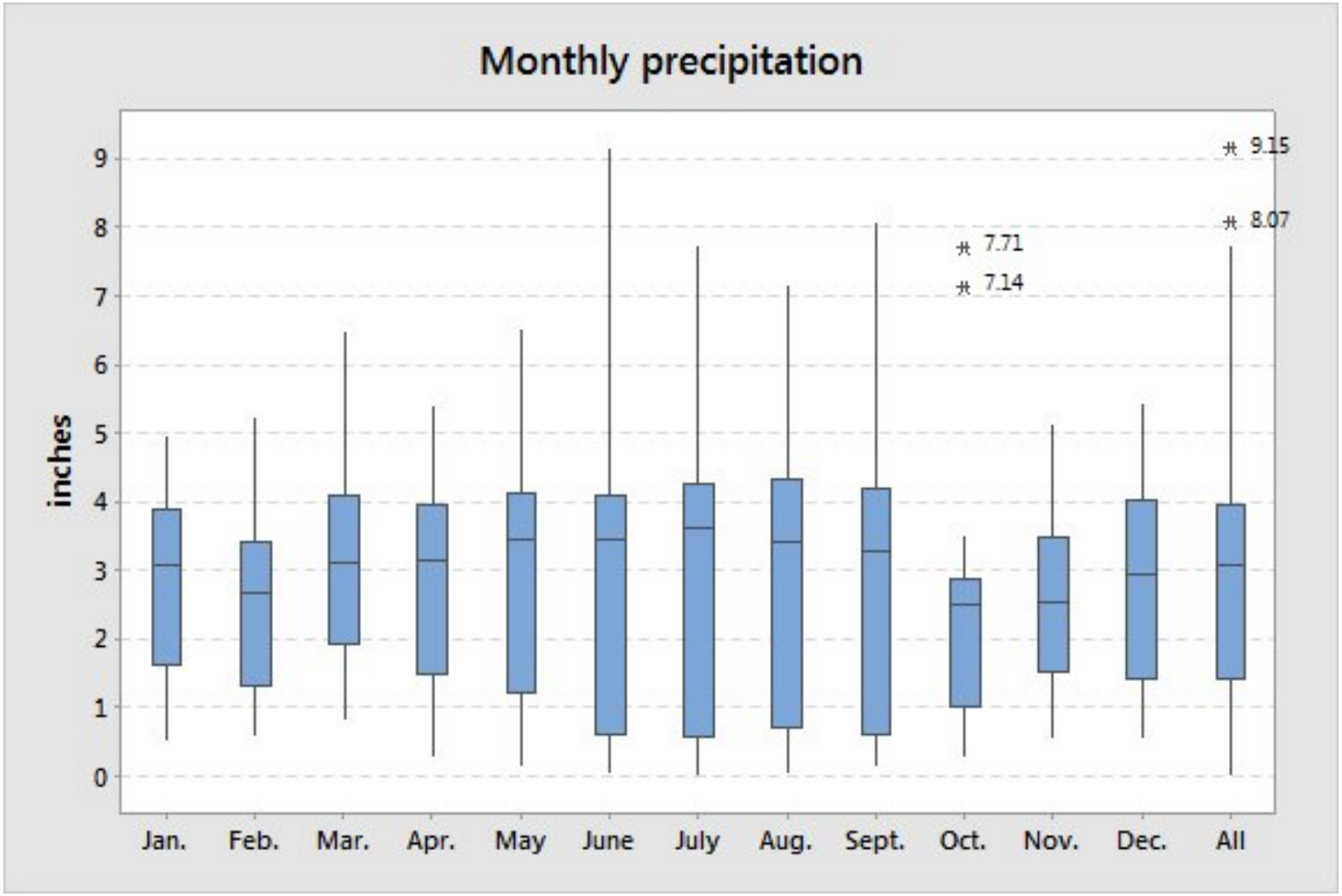
Split the data -categorical variables-

Normal Monthly and Annual Precipitation in Selected Cities

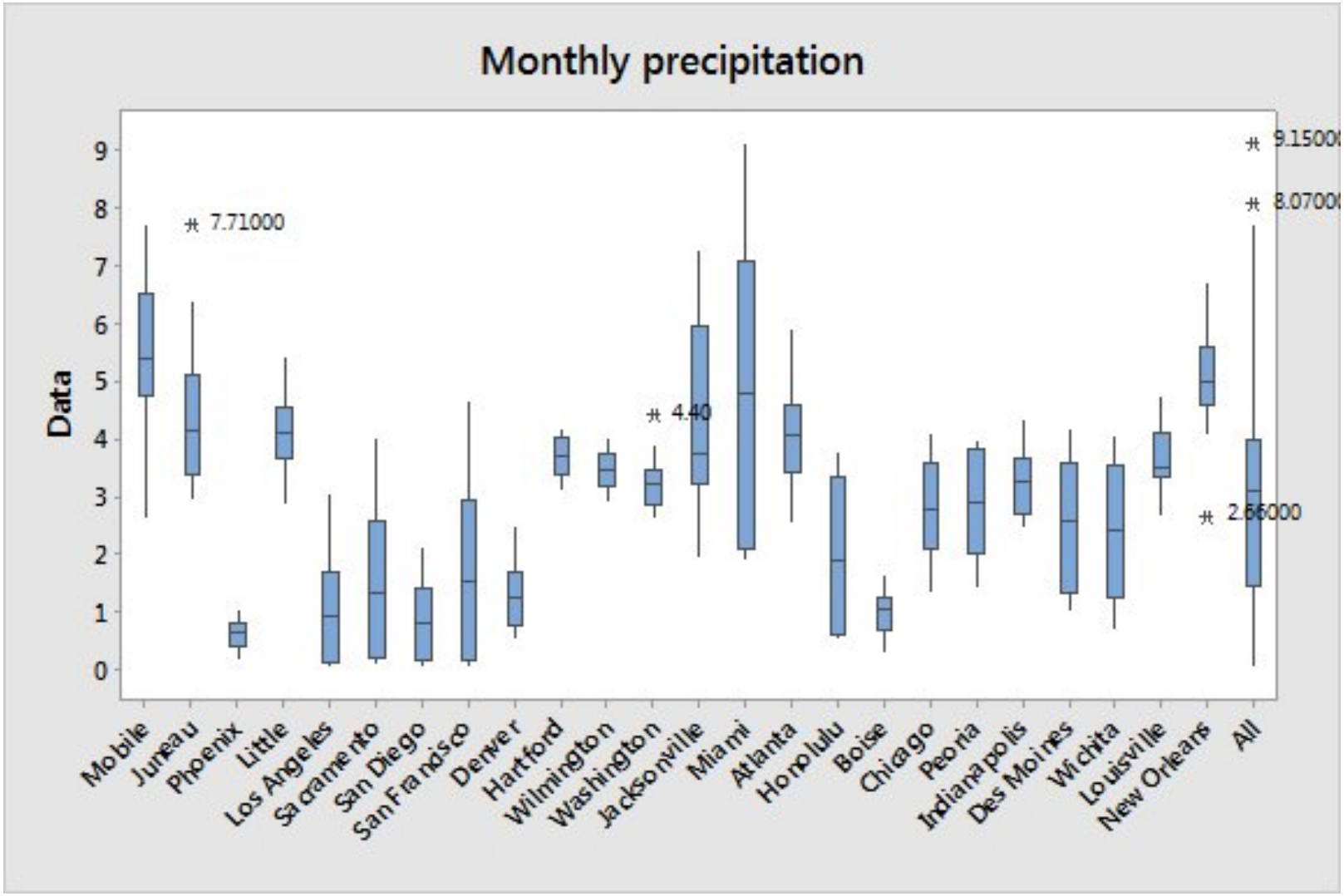
State	City	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
AL	Mobile	4.59	4.91	6.48	5.35	5.46	5.07	7.74	6.75	6.56	2.62	3.67	5.44
AK	Juneau	3.69	3.74	3.34	2.92	3.41	2.98	4.13	5.02	6.40	7.71	5.15	4.66
AZ	Phoenix	0.73	0.59	0.81	0.27	0.14	0.17	0.74	1.02	0.64	0.63	0.54	0.83
AR	Little Rock	3.91	3.83	4.69	5.41	5.29	3.67	3.63	3.07	4.26	2.84	4.37	4.23
CA	Los Angeles	3.06	2.49	1.76	0.93	0.14	0.04	0.01	0.10	0.15	0.26	1.52	1.62
	Sacramento	4.03	2.88	2.06	1.31	0.33	0.11	0.05	0.07	0.27	0.86	2.23	2.90
	San Diego	2.11	1.43	1.60	0.78	0.24	0.06	0.01	0.11	0.19	0.33	1.10	1.36
	San Francisco	4.65	3.23	2.64	1.53	0.32	0.11	0.03	0.05	0.19	1.06	2.35	3.55
CO	Denver	0.51	0.69	1.21	1.81	2.47	1.58	1.93	1.53	1.23	0.98	0.82	0.55
CT	Hartford	3.53	3.19	4.15	4.02	3.37	3.38	3.09	4.00	3.94	3.51	4.05	4.16
DE	Wilmington	3.11	2.99	3.87	3.39	3.23	3.51	3.90	4.03	3.59	2.89	3.33	3.54
DC	Washington	2.76	2.62	3.46	2.93	3.48	3.35	3.88	4.40	3.22	2.90	2.82	3.18
FL	Jacksonville	3.07	3.48	3.72	3.32	4.91	5.37	6.54	7.15	7.26	3.41	1.94	2.59
	Miami	2.08	2.05	1.89	3.07	6.53	9.15	5.98	7.02	8.07	7.14	2.71	1.86
GA	Atlanta	4.91	4.43	5.91	4.43	4.02	3.41	4.73	3.41	3.17	2.53	3.43	4.23
HI	Honolulu	3.79	2.72	3.48	1.49	1.21	0.49	0.54	0.60	0.62	1.88	3.22	3.43
ID	Boise	1.64	1.07	1.03	1.19	1.21	0.95	0.26	0.40	0.58	0.75	1.29	1.34
IL	Chicago	1.60	1.31	2.59	3.66	3.15	4.08	3.63	3.53	3.35	2.28	2.06	2.10
	Peoria	1.60	1.41	2.86	3.81	3.84	3.88	3.99	3.39	3.63	2.51	1.96	2.01
IN	Indianapolis	2.65	2.46	3.61	3.68	3.66	3.99	4.32	3.46	2.74	2.51	3.04	3.00
IA	Des Moines	1.01	1.12	2.20	3.21	3.96	4.18	3.22	4.11	3.09	2.16	1.52	1.05
KS	Wichita	0.68	0.85	2.01	2.30	3.91	4.06	3.62	2.80	3.45	2.47	1.47	0.99
KY	Louisville	3.38	3.23	4.73	4.11	4.15	3.60	4.10	3.31	3.35	2.63	3.49	3.48
LA	New Orleans	4.97	5.23	4.73	4.50	5.07	4.63	6.73	6.02	5.87	2.66	4.06	5.27

Source: U.S. National Oceanic and Atmospheric Administration, *Climatography of the United States*

Split the data -categorical variables-



Split the data -categorical variables-



Encoding variables –feature engineering-

Date of expiration

- days since agreement
- days until expiration
- separate month, year, and day of the week
- numeric day of the year

Creating new variables

- Combination of variables may be more useful than each individual variable

- The ratio of two predictors may prove more useful than using each one in a model