Statistics and Data Analysis, Fall 2015

Suggested Solution for Homework 1

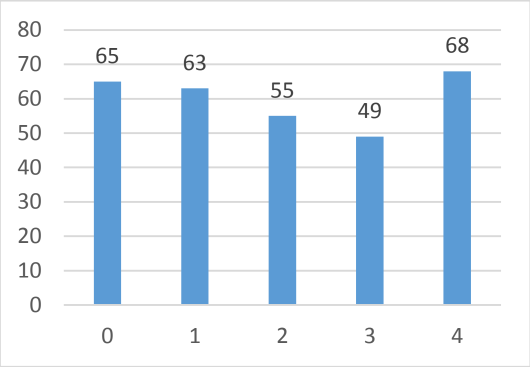
1.

(a) 592

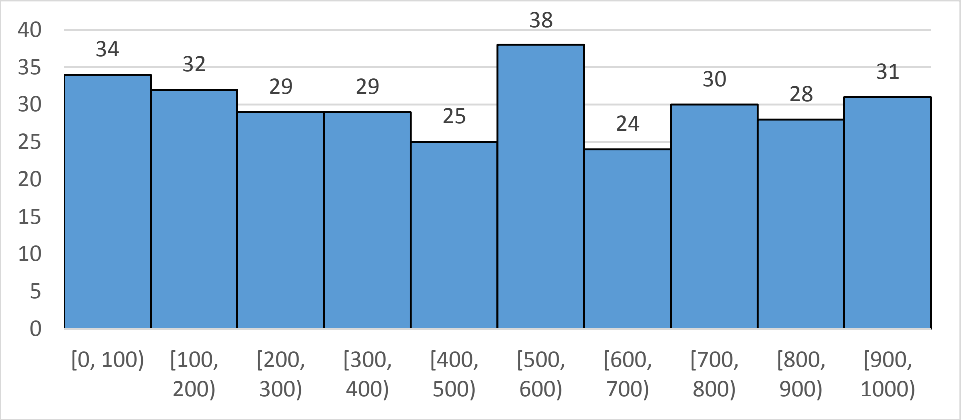
(b) 744.0529801

(c) 167

(d)



(e)

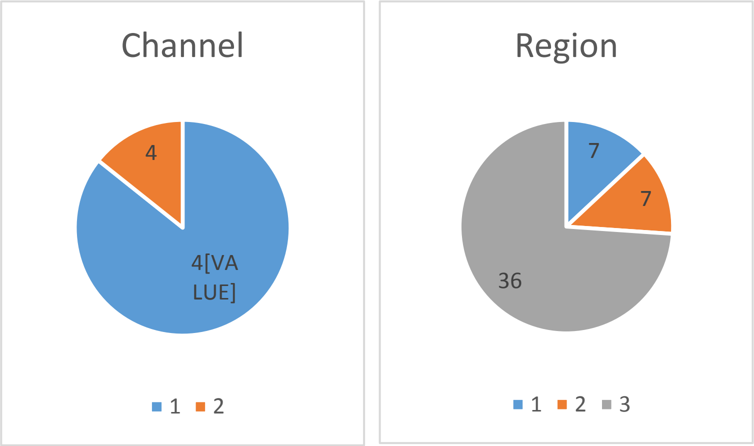


2.

(a)

Channel	1	2	3
1	18711	60869	36534
2	8321	11559	8132

(b)



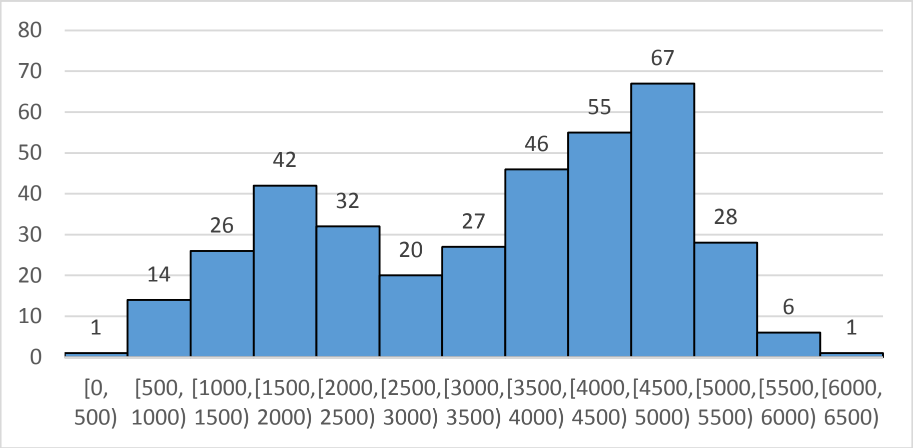
(c) 112

(d)

Channel	1	2	3	Total
1	9	2	37	48
2	10	9	45	64
Total	**19**	**11**	**82**	**112**

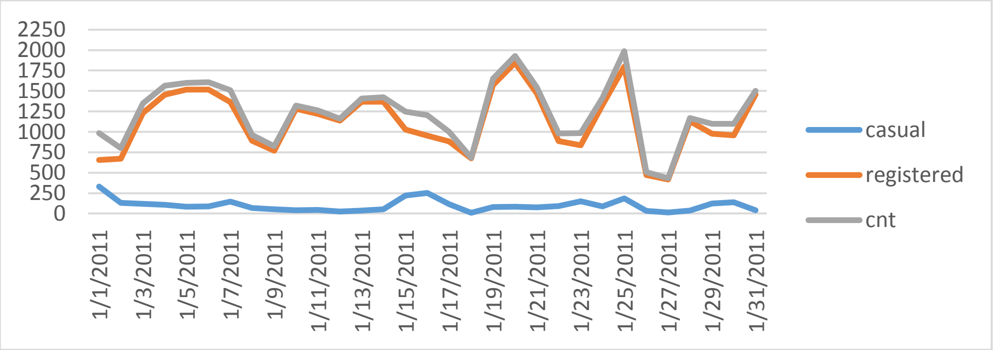
3.

(a)

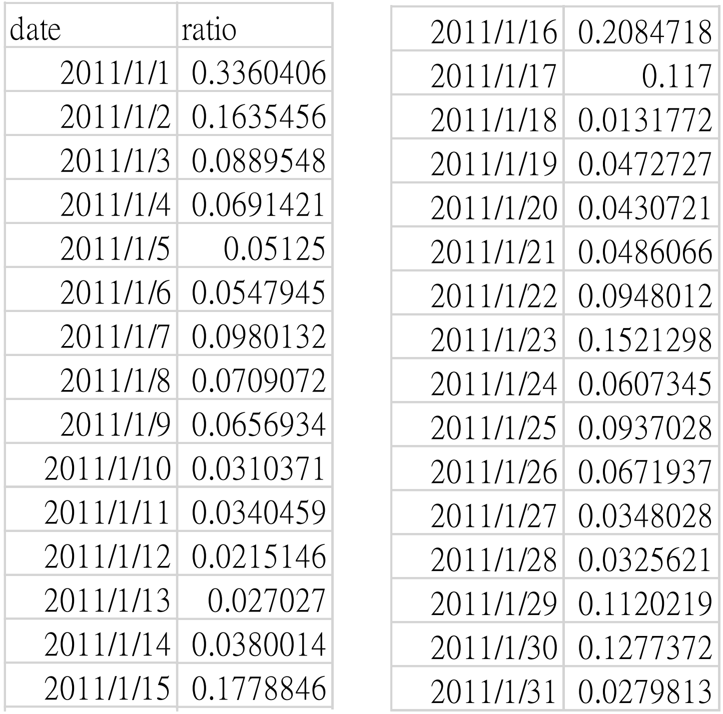


There are two peaks, and the highest one occurs at the interval [4000, 5000).

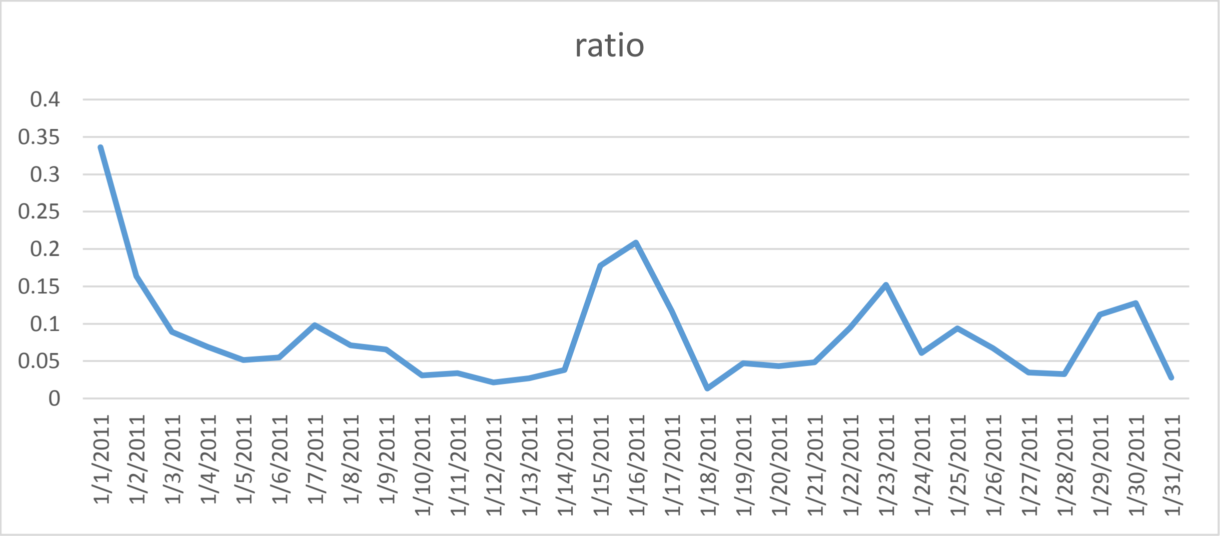
(b)



(c)



date	ratio
2011/1/1	0.3360406
2011/1/2	0.1635456
2011/1/3	0.0889548
2011/1/4	0.0691421
2011/1/5	0.05125
2011/1/6	0.0547945
2011/1/7	0.0980132
2011/1/8	0.0709072
2011/1/9	0.0656934
2011/1/10	0.0310371
2011/1/11	0.0340459
2011/1/12	0.0215146
2011/1/13	0.027027
2011/1/14	0.0380014
2011/1/15	0.1778846
2011/1/16	0.2084718
2011/1/17	0.117
2011/1/18	0.0131772
2011/1/19	0.0472727
2011/1/20	0.0430721
2011/1/21	0.0486066
2011/1/22	0.0948012
2011/1/23	0.1521298
2011/1/24	0.0607345
2011/1/25	0.0937028
2011/1/26	0.0671937
2011/1/27	0.0348028
2011/1/28	0.0325621
2011/1/29	0.1120219
2011/1/30	0.1277372
2011/1/31	0.0279813



It seems that the peaks occur at holidays like new year, and weekends.

(d) Total rentals = registered rentals + casual rentals

The ratio describes the proportion of casual rentals among all total rentals. Whether it's a "holiday" or a "non-workingday" might affect the number of casual rents, since there might be more visitors renting bikes. We know that the total rentals are higher on weekends than on weekdays, although registered rental increases on weekends as well, we can use the factor "holiday = 1" and "workingday = 0" with "weekday = 6 or 0" to predict the ratio of casual rentals to total rentals.

Please note that you cannot calculate the correlation coefficient between a qualitative variable and a quantitative variable!!!

4.

(a)

Mode of "*weathersit*" column equals to 1.

Median of "*weathersit*" column equals to 1.

Because "*weathersit*" column represents the weather situation —
(1 *for sunny or partly cloudy,* 2 *for misty and cloudy,* 3 *for light snow or light rain, and* 4 *for heavy snow or thunderstorm*), we consider it as nominal data. Mean of nominal data is meaningless.

(b) Frequency distribution:

Class	Humidity Frequency
[0, 10)	1
[10, 20)	1
[20, 30)	0
[30, 40)	10
[40, 50)	56
[50, 60)	74
[60, 70)	90
[70, 80)	77
[80, 90)	41
[90, 100]	15

(c) For "*humidity*" data, the distribution is said to be left-skewed, left tailed, or skewed to the left.

Mode	65	Largest
Median	64.75	Middle
Mean	64.36647671	Smallest

5.

(a) You may use either the formula for population or sample. If you want to study the daily bike rental in this particular year, you may use the formula for population; if you want to study daily bike rental in couple years, you may consider it as just a part of data, and you would use the formula for sample.

Using the formula for population:

Variance of "*casual*"	308587.5666
Standard deviation of "*casual*"	555.5065855

Using the formula for sample:

Variance of "*casual*"	309435.3346
Standard deviation of "*casual*"	556.2691207

(b)

Instant	*date*	*holiday*	*weekday*	*casual*	*registered*	*cnt*	*z-score*
149	2011/5/29	0	0	2355	2433	4788	3.016
185	2011/7/4	1	1	3065	2978	6043	4.292
197	2011/7/16	0	6	2418	3505	5923	3.129
247	2011/9/4	0	0	2521	2419	4940	3.314
282	2011/10/9	0	0	2397	3114	5511	3.091

(c) The answer depends on your investigation and explanation. If you do not consider them as outliers, you would explain what are possible factors which make these numbers so large or small. If you see them as outliers, you could not just say that it’s because its z-score is large or something like that, instead, give us reasons which support your idea.

6.

(a) Corr(temp, humidity) = 0.145776184
Corr(temp, cnt) = 0.771214198
Corr(humidity, cnt) = 0.001898085

(b) Temperature tends to affect the number of daily rentals more than humidity does. The correlation between temperature and the number of daily rentals is about 0.7712, and the correlation between humidity and the number of daily rentals is about 0.0019.

(c) We can see from the correlation coefficient that temperature has a stronger impact on the number of daily rentals than humidity.

You can then say if the result fit your intuition or not.