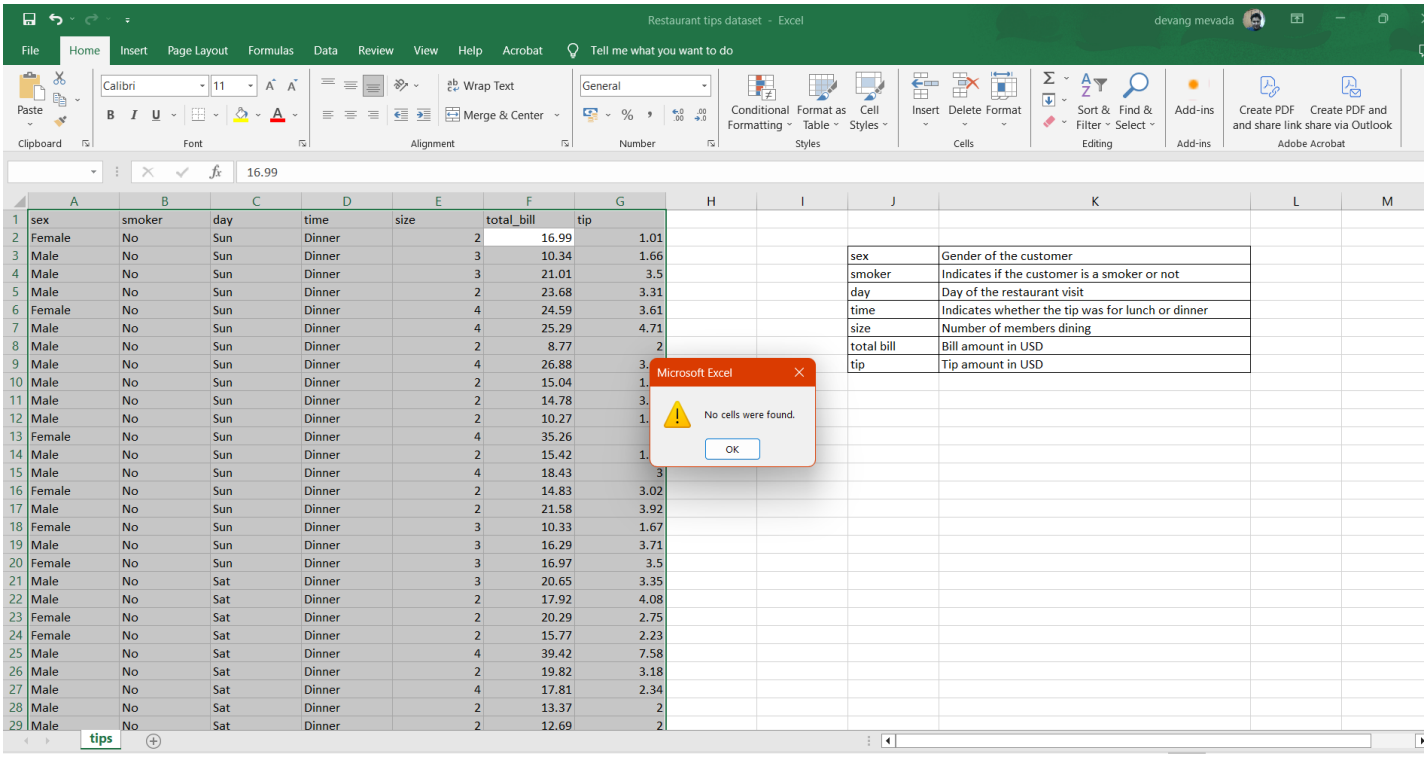
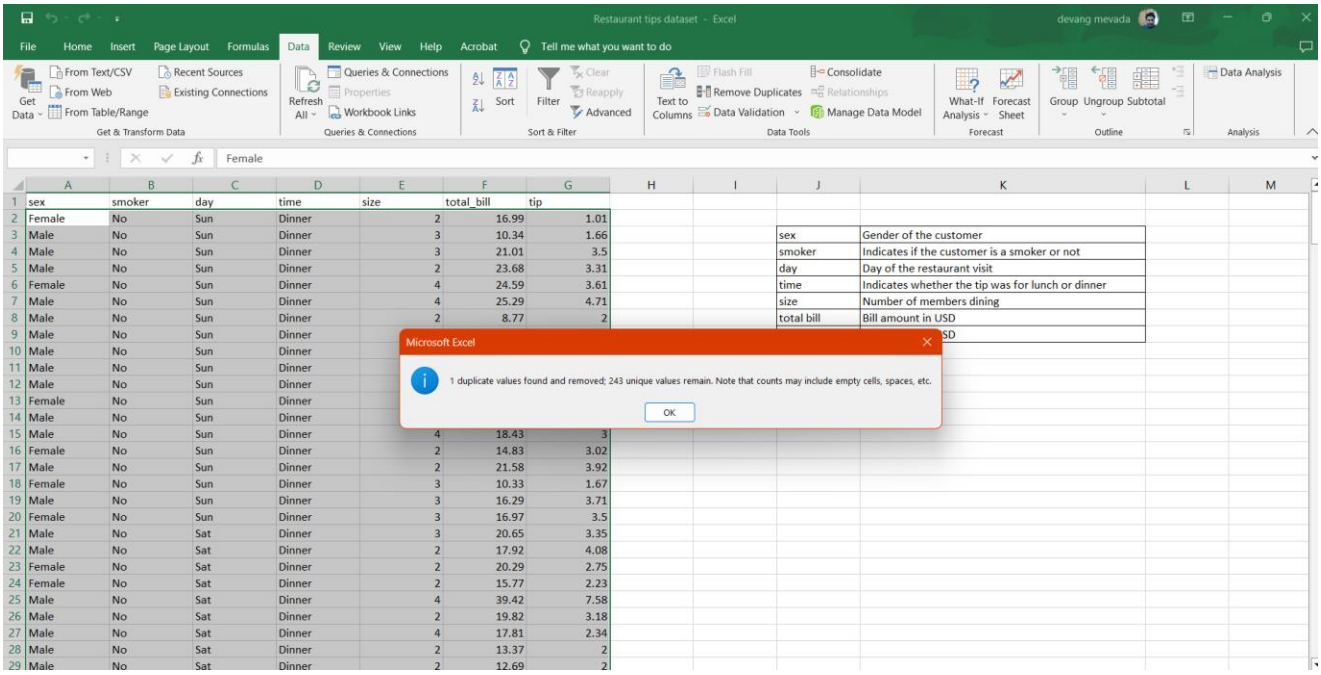
1. Find out if there are any missing values and clean the data

No missing values found



There is 1 duplicate values which I have removed

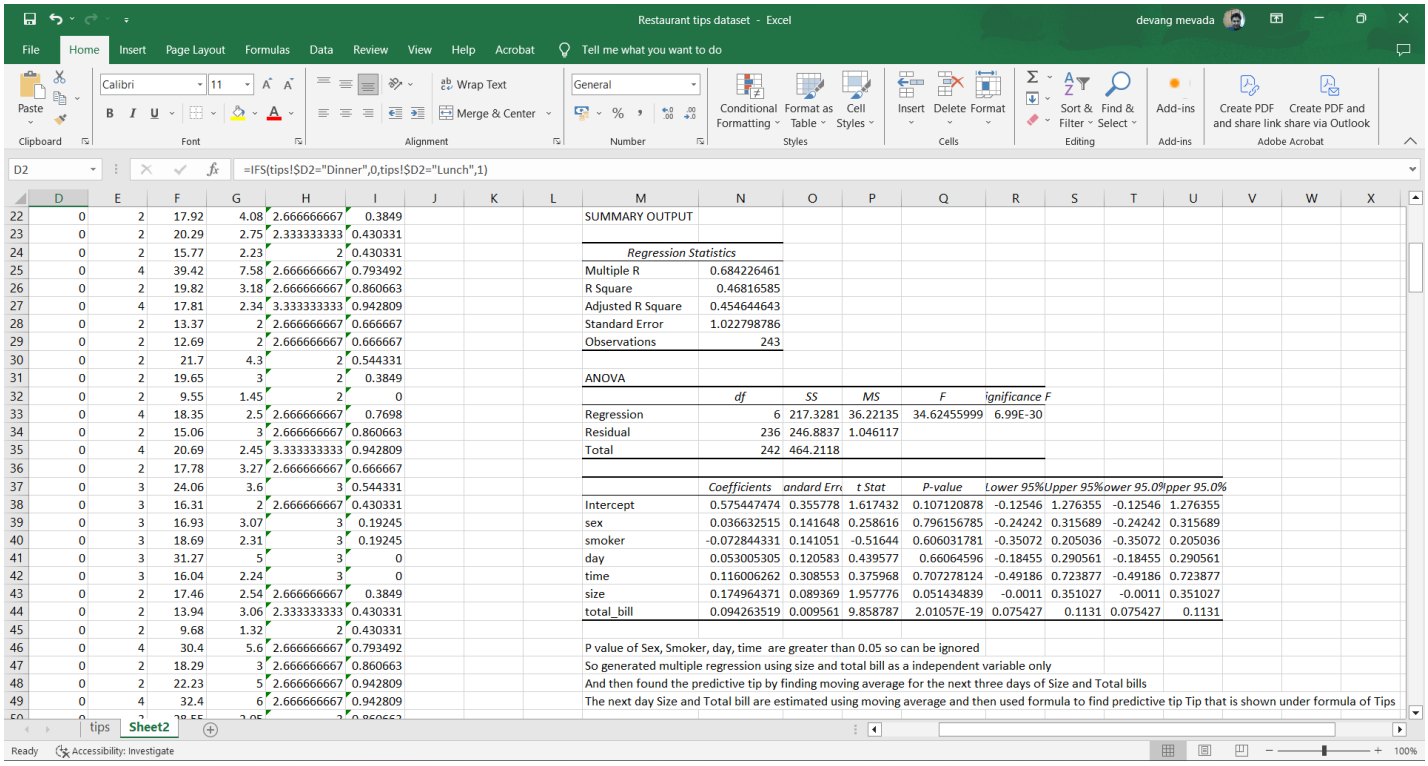


3. Find the features that are independent and dependent

   Independent features: Sex, Smoker, Date, Time, Size, Total bill

   Dependant feature: Tip

4. identify which predictive problem is needed
5. Encode the categorical variables to numeric values using IF conditions



Restaurant tips dataset - Excel

=IFS(tips!$D2="Dinner",0,tips!$D2="Lunch",1)

	D	E	F	G	H	I
22	0	2	17.92	4.08	2.666666667	0.3849
23	0	2	20.29	2.75	2.333333333	0.430331
24	0	2	15.77	2.23	2	0.430331
25	0	4	39.42	7.58	2.666666667	0.793492
26	0	2	19.82	3.18	2.666666667	0.860663
27	0	4	17.81	2.34	3.333333333	0.942809
28	0	2	13.37	2	2.666666667	0.666667
29	0	2	12.69	2	2.666666667	0.666667
30	0	2	21.7	4.3	2	0.544331
31	0	2	19.65	3	2	0.3849
32	0	2	9.55	1.45	2	0
33	0	4	18.35	2.5	2.666666667	0.7698
34	0	2	15.06	3	2.666666667	0.860663
35	0	4	20.69	2.45	3.333333333	0.942809
36	0	2	17.78	3.27	2.666666667	0.666667
37	0	3	24.06	3.6	3	0.544331
38	0	3	16.31	2	2.666666667	0.430331
39	0	3	16.93	3.07	3	0.19245
40	0	3	18.69	2.31	3	0.19245
41	0	3	31.27	5	3	0
42	0	3	16.04	2.24	3	0
43	0	2	17.46	2.54	2.666666667	0.3849
44	0	2	13.94	3.06	2.333333333	0.430331
45	0	2	9.68	1.32	2	0.430331
46	0	4	30.4	5.6	2.666666667	0.793492
47	0	2	18.29	3	2.666666667	0.860663
48	0	2	22.23	5	2.666666667	0.942809
49	0	4	32.4	6	2.666666667	0.942809

	M	N	O	P	Q	R	S	T	U
22	SUMMARY OUTPUT								
23									
24	*Regression Statistics*								
25	Multiple R	0.684226461							
26	R Square	0.46816585							
27	Adjusted R Square	0.454644643							
28	Standard Error	1.022798786							
29	Observations	243							
30									
31	ANOVA								
32		*df*	*SS*	*MS*	*F*	*ignificance F*			
33	Regression	6	217.3281	36.22135	34.62455999	6.99E-30			
34	Residual	236	246.8837	1.046117					
35	Total	242	464.2118						
36									
37		*Coefficients*	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
38	Intercept	0.575447474	0.355778	1.617432	0.107120878	-0.12546	1.276355	-0.12546	1.276355
39	sex	0.036632515	0.141648	0.258616	0.796156785	-0.24242	0.315689	-0.24242	0.315689
40	smoker	-0.072844331	0.141051	-0.51644	0.606031781	-0.35072	0.205036	-0.35072	0.205036
41	day	0.053005305	0.120583	0.439577	0.66064596	-0.18455	0.290561	-0.18455	0.290561
42	time	0.116006262	0.308553	0.375968	0.707278124	-0.49186	0.723877	-0.49186	0.723877
43	size	0.174964371	0.089369	1.957776	0.051434839	-0.0011	0.351027	-0.0011	0.351027
44	total_bill	0.094263519	0.009561	9.858787	2.01057E-19	0.075427	0.1131	0.075427	0.1131
45									
46	P value of Sex, Smoker, day, time are greater than 0.05 so can be ignored								
47	So generated multiple regression using size and total bill as a independent variable only								
48	And then found the predictive tip by finding moving average for the next three days of Size and Total bills								
49	The next day Size and Total bill are estimated using moving average and then used formula to find predictive tip Tip that is shown under formula of Tips								

tips | Sheet2

Ready | Accessibility: Investigate

6.Build an appropriate model with the dataset

7.Calculate the predicted and actual tips values

Restaurant tips dataset - Excel

devang mevada

File Home Insert Page Layout Formulas Data Review View Help Acrobat Tell me what you want to do

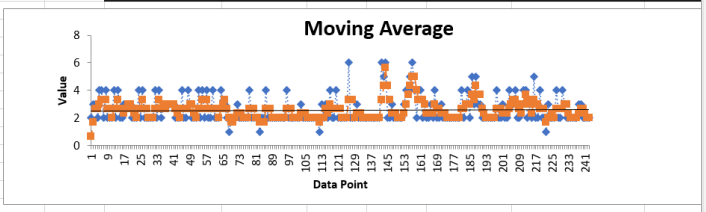
Get Data | From Text/CSV | From Web | From Table/Range | Recent Sources | Existing Connections | Refresh All | Queries & Connections | Properties | Workbook Links | Sort | Filter | Clear | Reapply | Advanced | Text to Columns | Flash Fill | Remove Duplicates | Data Validation | Consolidate | Relationships | Manage Data Model | What-If Analysis | Forecast Sheet | Group | Ungroup | Subtotal | Data Analysis

Get & Transform Data | Queries & Connections | Sort & Filter | Data Tools | Forecast | Outline | Analysis

K3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	sex	smoker	day	time	size	total_bill	tip	Predictive Tip	Error										
2	1	0	3	0	2	16.99	1.01	0.666666667	0.7698										
3	0	0	3	0	3	10.34	1.66	1.666666667	1.539601										
4	0	0	3	0	3	21.01	3.5	2.666666667	1.105542										
5	0	0	3	0	2	23.68	3.31	2.666666667	0.881917										
6	1	0	3	0	4	24.59	3.61	3	0.720082										
7	0	0	3	0	4	25.29	4.71	3.333333333	0.793492										
8	0	0	3	0	2	8.77	2	3.333333333	1.036375										
9	0	0	3	0	4	26.88	3.12	3.333333333	0.942809										
10	0	0	3	0	2	15.04	1.96	2.666666667	0.942809										
11	0	0	3	0	2	14.78	3.23	2.666666667	0.666667										
12	0	0	3	0	2	10.27	1.71	2	0.544331										
13	1	0	3	0	4	35.26	5	2.666666667	0.860663										
14	0	0	3	0	2	15.42	1.57	2.666666667	0.860663										
15	0	0	3	0	4	18.43	3	3.333333333	0.942809										
16	1	0	3	0	2	14.83	3.02	2.666666667	0.666667										
17	0	0	3	0	2	21.58	3.92	2.666666667	0.666667										
18	1	0	3	0	3	10.33	1.67	2.333333333	0.666667										
19	0	0	3	0	3	16.29	3.71	2.666666667	0.57735										
20	1	0	3	0	4	16.97	3.5	3	0.430331										
21	0	0	2	0	3	20.65	3.35	3	0.19245										
22	0	0	2	0	2	17.92	4.08	2.666666667	0.3849										
23	1	0	2	0	2	20.29	2.75	2.333333333	0.430331										
24	1	0	2	0	2	15.77	2.23	2	0.430331										
25	0	0	2	0	4	39.42	7.58	2.666666667	0.793492										
26	0	0	2	0	2	19.82	3.18	2.666666667	0.860663										
27	0	0	2	0	4	17.81	2.34	3.333333333	0.942809										
28	0	0	2	0	2	13.37	2	2.666666667	0.666667										
29	0	0	2	0	2	12.69	2	2.666666667	0.666667										

	sex	smoker	day	time	size	total_bill
sex	1					
smoker	-0.009930188	1				
day	-0.224387598	-0.02501	1			
time	0.198128623	-0.06391	-0.87313	1		
size	-0.083248017	-0.13056	0.162525	-0.100045303	1	
total_bill	-0.141349744	0.090136	0.169978	-0.179231854	0.597588931	1
tip	-0.085273975	0.009763	0.131798	-0.11759639	0.488400395	0.674998



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.684226461
R Square	0.46816585
Adjusted R Square	0.454644643
Standard Error	1.022798786
Observations	243

Sum Square	129.81893
Count	243
RMSE	0.53423428

Restaurant tips dataset - Excel

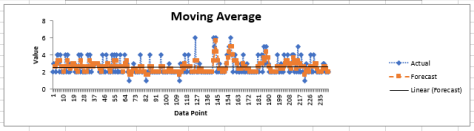
devang mevada

File Home Insert Page Layout Formulas Data Review View Help Acrobat Tell me what you want to do

Paste | Calibri | 11 | Wrap Text | Merge & Center | General | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | Sort & Filter | Find & Select | Add-ins | Create PDF and share link | Create PDF and share via Outlook

Clipboard | Font | Alignment | Number | Styles | Cells | Editing | Add-ins | Adobe Acrobat

R28



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.68422646
R Square	0.46816585
Adjusted R Squar	0.45464464
Standard Error	1.02279879
Observations	243

ANOVA

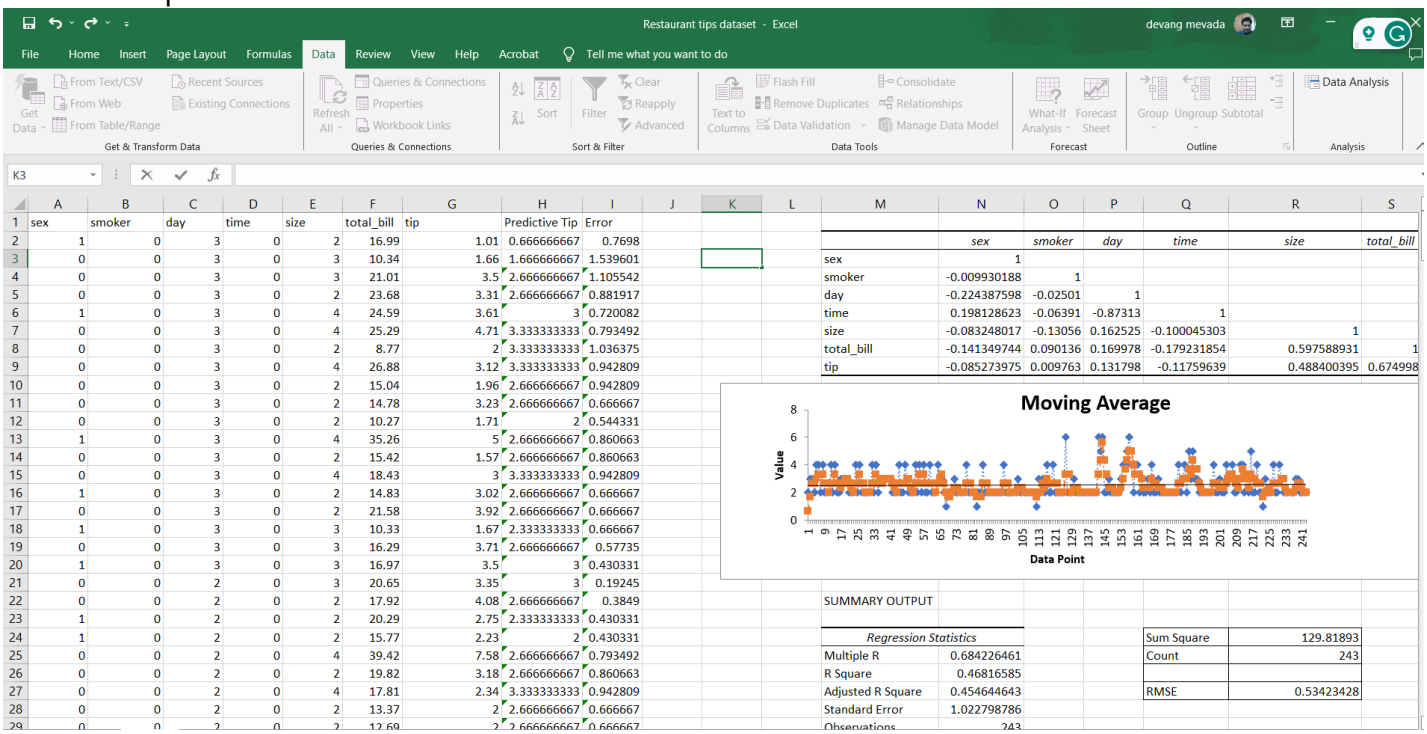
	df	SS	MS	F	Significance F
Regression	6	217.328	36.2214	34.62456	7E-30
Residual	236	246.884	1.04612		
Total	242	464.212			

	Coefficients	Standard Er	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.57544747	0.35578	1.61743	0.10712088	-0.1255	1.27635	-0.1255	1.27635
sex	0.03663251	0.14165	0.25862	0.79615679	-0.2424	0.31569	-0.2424	0.31569
smoker	-0.07284433	0.14105	-0.5164	0.60603178	-0.3507	0.20504	-0.3507	0.20504
day	0.0530053	0.12058	0.43958	0.66064596	-0.1846	0.29056	-0.1846	0.29056
time	0.11600626	0.30855	0.37597	0.70727812	-0.4919	0.72388	-0.4919	0.72388
size	0.17496437	0.08937	1.95778	0.05143484	-0.0011	0.35103	-0.0011	0.35103
total_bill	0.09426352	0.00956	9.85879	2.01066E-19	0.07543	0.1131	0.07543	0.1131

P value of Sex, Smoker, day, time are greater than 0.05 so can be ignored

So generated multiple regression using size and total bill as a independent variable only

8. Calculate the RMSE (Root Mean Square Error) of the model. RMSE is the root of the mean of square errors



	A	B	C	D	E	F	G	H	I
1	sex	smoker	day	time	size	total_bill	tip	Predictive Tip	Error
2	1	0	3	0	2	16.99	1.01	0.666666667	0.7698
3	0	0	3	0	3	10.34	1.66	1.666666667	1.539601
4	0	0	3	0	3	21.01	3.5	2.666666667	1.105542
5	0	0	3	0	2	23.68	3.31	2.666666667	0.881917
6	1	0	3	0	4	24.59	3.61	3	0.720082
7	0	0	3	0	4	25.29	4.71	3.333333333	0.793492
8	0	0	3	0	2	8.77	2	3.333333333	1.036375
9	0	0	3	0	4	26.88	3.12	3.333333333	0.942809
10	0	0	3	0	2	15.04	1.96	2.666666667	0.942809
11	0	0	3	0	2	14.78	3.23	2.666666667	0.666667
12	0	0	3	0	2	10.27	1.71	2	0.544331
13	1	0	3	0	4	35.26	5	2.666666667	0.860663
14	0	0	3	0	2	15.42	1.57	2.666666667	0.860663
15	0	0	3	0	4	18.43	3	3.333333333	0.942809
16	1	0	3	0	2	14.83	3.02	2.666666667	0.666667
17	0	0	3	0	2	21.58	3.92	2.666666667	0.666667
18	1	0	3	0	3	10.33	1.67	2.333333333	0.666667
19	0	0	3	0	3	16.29	3.71	2.666666667	0.57735
20	1	0	3	0	3	16.97	3.5	3	0.430331
21	0	0	2	0	3	20.65	3.35	3	0.19245
22	0	0	2	0	2	17.92	4.08	2.666666667	0.3849
23	1	0	2	0	2	20.29	2.75	2.333333333	0.430331
24	1	0	2	0	2	15.77	2.23	2	0.430331
25	0	0	2	0	4	39.42	7.58	2.666666667	0.793492
26	0	0	2	0	2	19.82	3.18	2.666666667	0.860663
27	0	0	2	0	4	17.81	2.34	3.333333333	0.942809
28	0	0	2	0	2	13.37	2	2.666666667	0.666667
29	0	0	2	0	2	12.69	2	2.666666667	0.666667

	sex	smoker	day	time	size	total_bill
sex	1					
smoker	-0.009930188	1				
day	-0.224387598	-0.02501	1			
time	0.198128623	-0.06391	-0.87313	1		
size	-0.083248017	-0.13056	0.162525	-0.100045303	1	
total_bill	-0.141349744	0.090136	0.169978	-0.179231854	0.597588931	1
tip	-0.085273975	0.009763	0.131798	-0.11759639	0.488400395	0.674998

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.684226461
R Square	0.46816585
Adjusted R Square	0.454644643
Standard Error	1.022798786
Observations	243

Sum Square	129.81893
Count	243
RMSE	0.53423428