Predicting restaurant tips using predictive analytics on Excel

DESCRIPTION: The dataset in file Restaurant tips dataset.xlsx contains tips data for different customers. The following are the features in the dataset:

Sex - Gender of the customer
Smoker- Indicates if the customer is a smoker or not
Day - Day of the restaurant visit
Time - Indicates whether the tip was for lunch or dinner
Size - Number of members dining
Total bill - Bill amount in USD
Tip - Tip amount in USD

TOOLS USED: Microsoft Excel, Data Analysis Add-in.

STEPS FOR EXECUTION:

Exploratory Data Analysis:

- **Data Cleaning** - Missing entries were removed from the dataset. Duplicate and redundant entries were filtered and removed.
- **Feature Identification** - The following features were found:
  - Independent Features - sex, smoker, day, time, size, total bill
  - Dependent Features – tip
- **Feature Encoding** – The following categorical variables were encoded to numeric values using IF conditions:
  - Sex: Female - 1, Male - 2
  - Smoker: No - 3, Yes- 4
  - Day: Sun - 5, Sat - 6, Fri - 7, Thur - 8
  - Time: Lunch - 10, Dinner – 11
- **Standardization in excel using STANDARDIZE()** – All features were standardized using the excel STANDARDIZE Z-Score function.

- **Feature Analysis –** Relation between features was determined using the Correlation and Covariance matrix. We select the independent features that affect the dependent feature tip the most.
    - Feature smoker(Y/N) shows 0 correlation and covariance with tip. Approximate values were Correlation ~ 0.009 and covariance ~0.006.
    - Feature size and total bill show high positive correlation with tip. Approximate values were Correlation ~ 0.48 and covariance ~0.64 for features size and tip, correlation ~ 0.67 and covariance ~8.29 for features total bill and tip.
    - So, we can remove the feature smoker from our model as it has negligible impact on the dependent variable tip. The features size and total bill have high impact on the feature tip.

Multiple Linear Regression model was applied on the dataset to predict the restaurant tips. The predicted tips range from 1 to 10. The predictive models were built and applied on the given dataset.

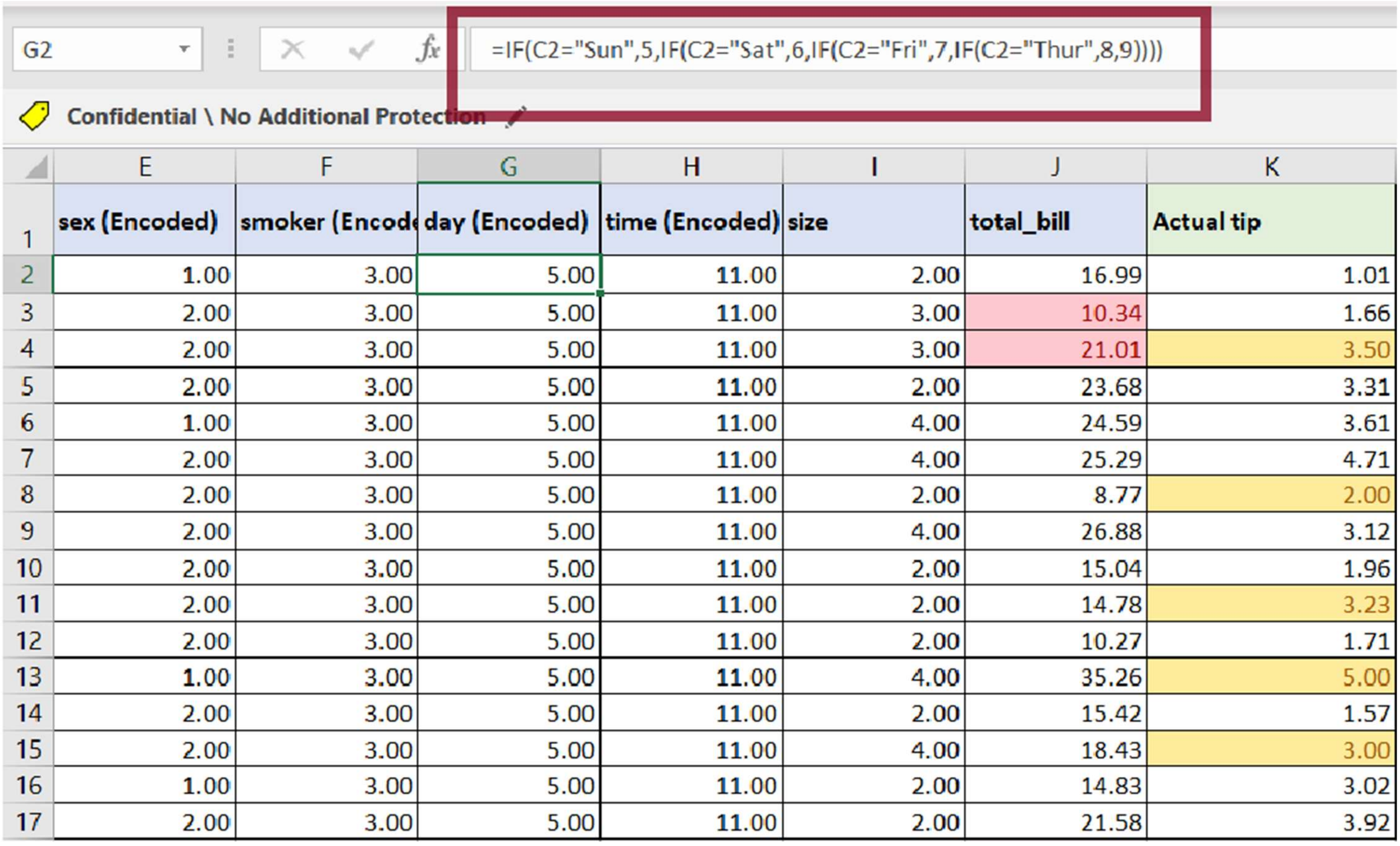
- Regression Model 1 - model trained by using all the non-standardized independent Features - sex, smoker, day, time, size, total bill. The P-value of all the features except size, total bill was above 0.5 tolerance level. Thus, the other features include randomness in the model and can be ignored.
- Regression Model 2- model trained by using the non-standardized independent Features - size, total bill. The P-value of these features was below 0.5 tolerance level.
- Regression Model 3 - model trained by using all the standardized independent Features - sex, smoker, day, time, size, total bill. The P-value of all the features except size, total bill was above 0.5 tolerance level. Thus, the other features include randomness in the model and can be ignored. R-squared index was similar to Model 1.
- Regression Model 4- model trained by using the standardized independent Features - size, total bill. The P-value of these features was below 0.5 tolerance level. R-Squared index was similar to Model 2.

• Predicted tip values using the above regression models were compared to the actual values.

• **RMSE (Root Mean Square Error)** of the model was calculated. RMSE is root of mean of square errors. Following was observed:

- RMSE Regression Model 1 - 1.0079
- RMSE Regression Model 2 - 1.0091
- Regression Model 2 has a better RMSE than Regression Model 1

Feature Encoding:

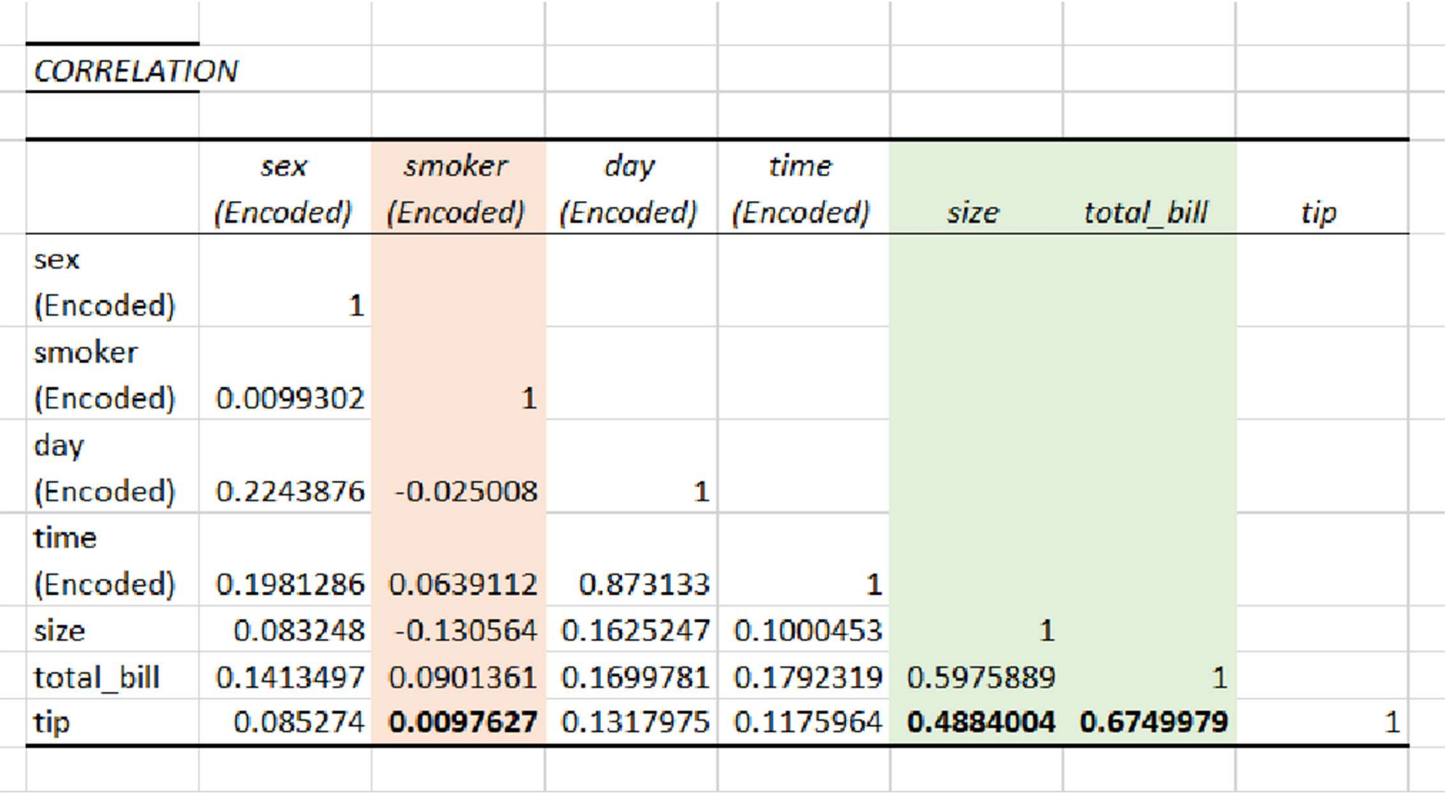


G2 =IF(C2="Sun",5,IF(C2="Sat",6,IF(C2="Fri",7,IF(C2="Thur",8,9))))

Confidential \ No Additional Protection

	E	F	G	H	I	J	K
1	sex (Encoded)	smoker (Encode	day (Encoded)	time (Encoded)	size	total_bill	Actual tip
2	1.00	3.00	5.00	11.00	2.00	16.99	1.01
3	2.00	3.00	5.00	11.00	3.00	10.34	1.66
4	2.00	3.00	5.00	11.00	3.00	21.01	3.50
5	2.00	3.00	5.00	11.00	2.00	23.68	3.31
6	1.00	3.00	5.00	11.00	4.00	24.59	3.61
7	2.00	3.00	5.00	11.00	4.00	25.29	4.71
8	2.00	3.00	5.00	11.00	2.00	8.77	2.00
9	2.00	3.00	5.00	11.00	4.00	26.88	3.12
10	2.00	3.00	5.00	11.00	2.00	15.04	1.96
11	2.00	3.00	5.00	11.00	2.00	14.78	3.23
12	2.00	3.00	5.00	11.00	2.00	10.27	1.71
13	1.00	3.00	5.00	11.00	4.00	35.26	5.00
14	2.00	3.00	5.00	11.00	2.00	15.42	1.57
15	2.00	3.00	5.00	11.00	4.00	18.43	3.00
16	1.00	3.00	5.00	11.00	2.00	14.83	3.02
17	2.00	3.00	5.00	11.00	2.00	21.58	3.92

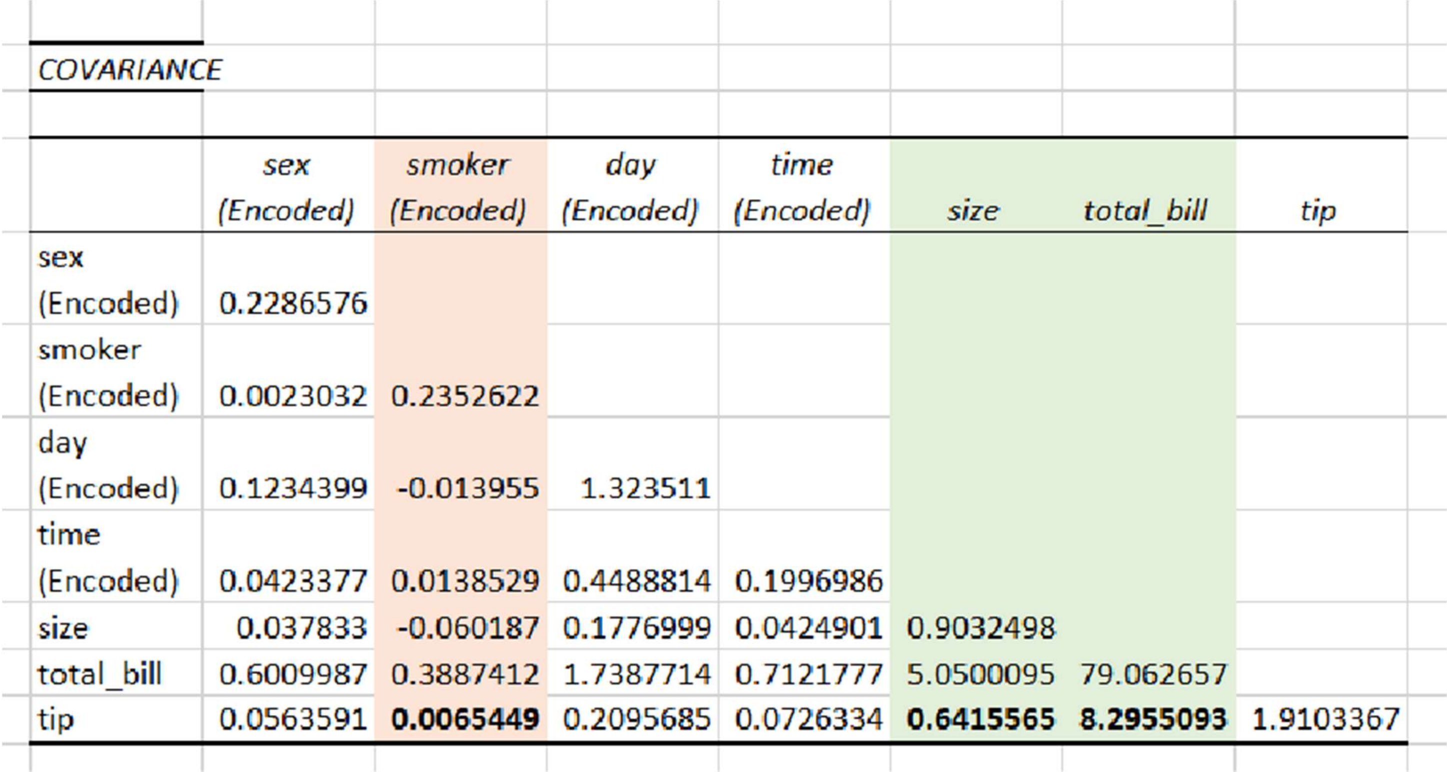
Feature Analysis:



CORRELATION

	sex (Encoded)	*smoker (Encoded)*	*day (Encoded)*	*time (Encoded)*	*size*	*total_bill*	*tip*
sex (Encoded)	1						
smoker (Encoded)	0.0099302	1					
day (Encoded)	0.2243876	-0.025008	1				
time (Encoded)	0.1981286	0.0639112	0.873133	1			
size	0.083248	-0.130564	0.1625247	0.1000453	1		
total_bill	0.1413497	0.0901361	0.1699781	0.1792319	0.5975889	1	
tip	0.085274	**0.0097627**	0.1317975	0.1175964	**0.4884004**	**0.6749979**	1

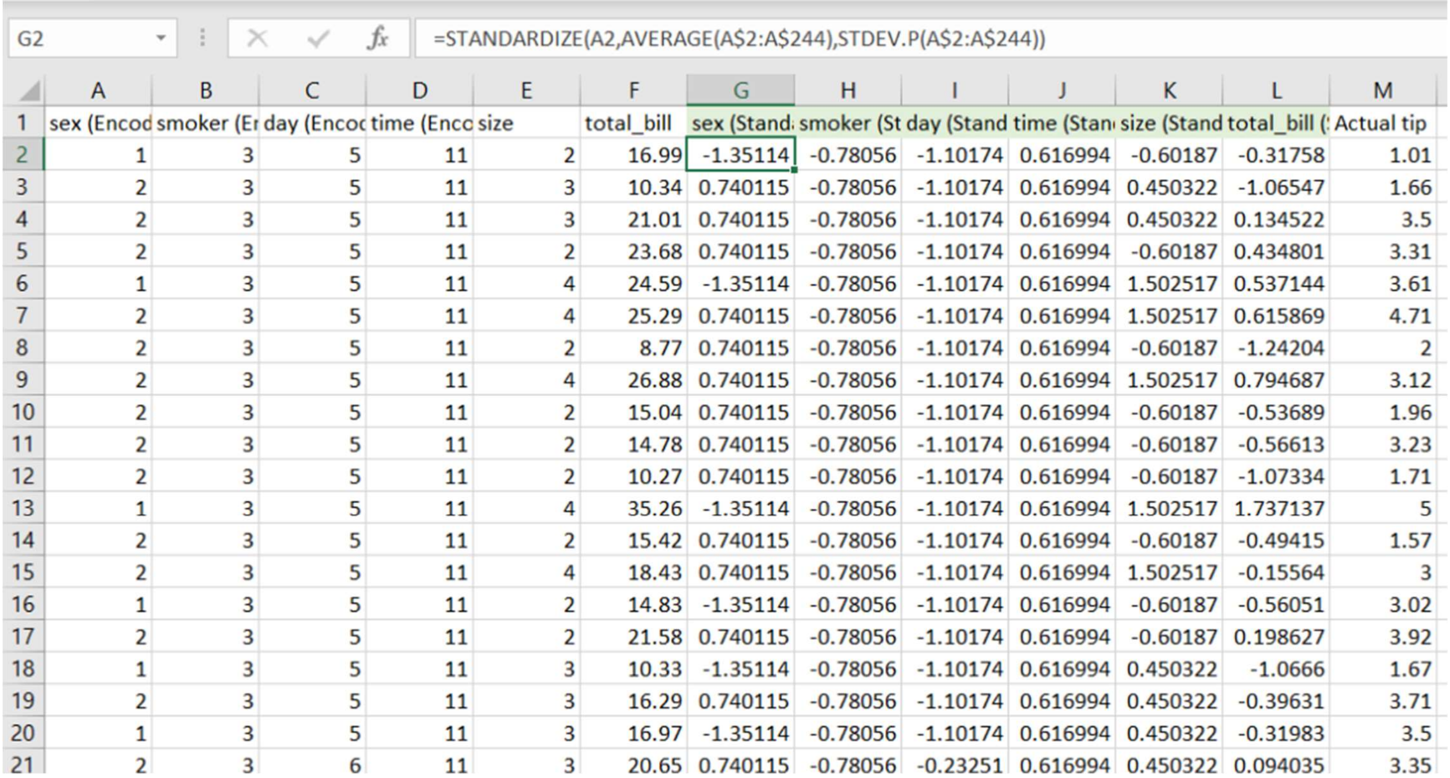
FEATURE STANDARDIZATION:



COVARIANCE

	sex (Encoded)	*smoker (Encoded)*	*day (Encoded)*	*time (Encoded)*	*size*	*total_bill*	*tip*
sex (Encoded)	0.2286576						
smoker (Encoded)	0.0023032	0.2352622					
day (Encoded)	0.1234399	-0.013955	1.323511				
time (Encoded)	0.0423377	0.0138529	0.4488814	0.1996986			
size	0.037833	-0.060187	0.1776999	0.0424901	0.9032498		
total_bill	0.6009987	0.3887412	1.7387714	0.7121777	5.0500095	79.062657	
tip	0.0563591	**0.0065449**	0.2095685	0.0726334	**0.6415565**	**8.2955093**	1.9103367

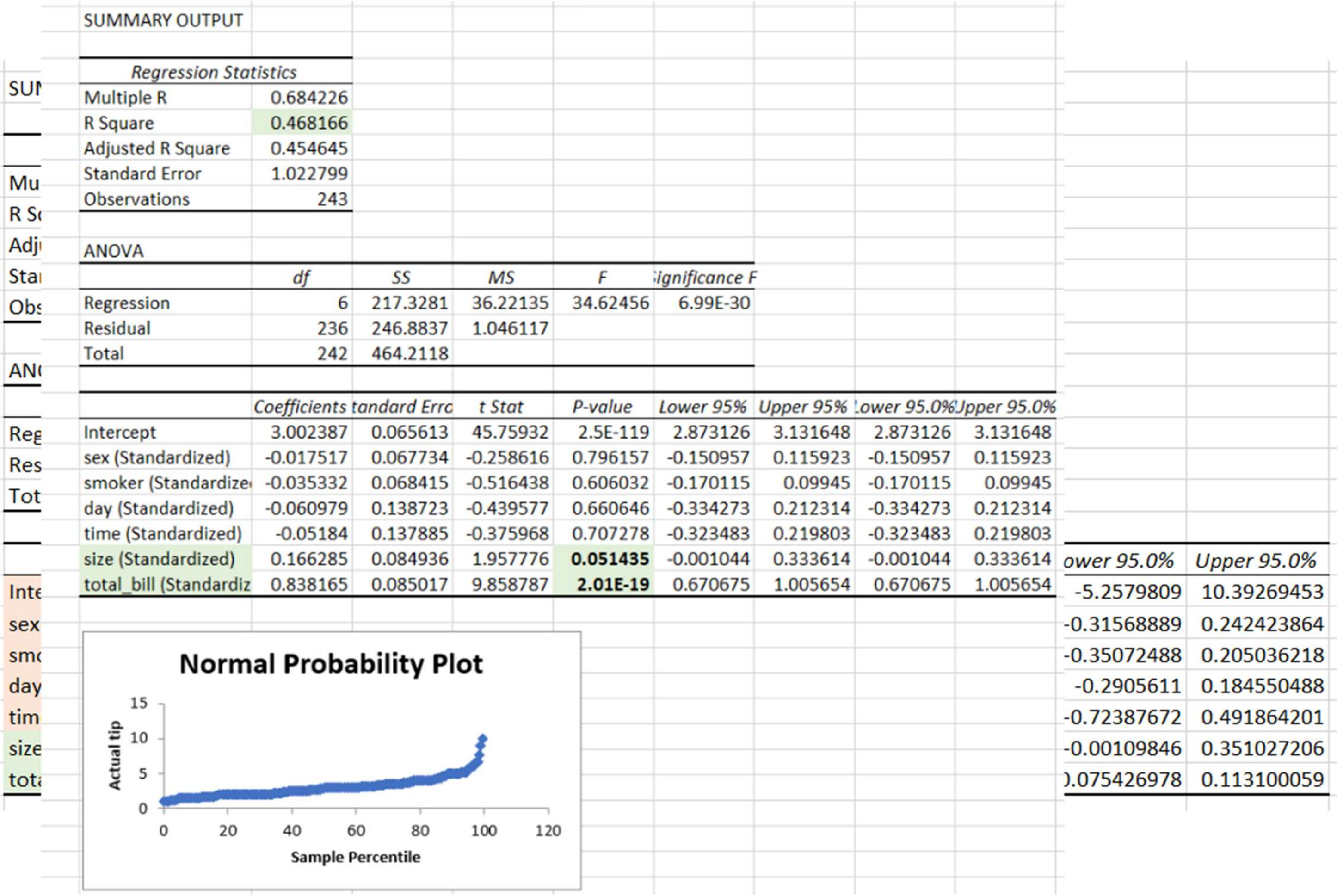
Multiple Linear Regression Model 1:



=STANDARDIZE(A2,AVERAGE(A$2:A$244),STDEV.P(A$2:A$244))

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	sex (Encod	smoker (Er	day (Encoc	time (Enco	size	total_bill	sex (Stand	smoker (St	day (Stand	time (Stan	size (Stand	total_bill (:	Actual tip
2	1	3	5	11	2	16.99	-1.35114	-0.78056	-1.10174	0.616994	-0.60187	-0.31758	1.01
3	2	3	5	11	3	10.34	0.740115	-0.78056	-1.10174	0.616994	0.450322	-1.06547	1.66
4	2	3	5	11	3	21.01	0.740115	-0.78056	-1.10174	0.616994	0.450322	0.134522	3.5
5	2	3	5	11	2	23.68	0.740115	-0.78056	-1.10174	0.616994	-0.60187	0.434801	3.31
6	1	3	5	11	4	24.59	-1.35114	-0.78056	-1.10174	0.616994	1.502517	0.537144	3.61
7	2	3	5	11	4	25.29	0.740115	-0.78056	-1.10174	0.616994	1.502517	0.615869	4.71
8	2	3	5	11	2	8.77	0.740115	-0.78056	-1.10174	0.616994	-0.60187	-1.24204	2
9	2	3	5	11	4	26.88	0.740115	-0.78056	-1.10174	0.616994	1.502517	0.794687	3.12
10	2	3	5	11	2	15.04	0.740115	-0.78056	-1.10174	0.616994	-0.60187	-0.53689	1.96
11	2	3	5	11	2	14.78	0.740115	-0.78056	-1.10174	0.616994	-0.60187	-0.56613	3.23
12	2	3	5	11	2	10.27	0.740115	-0.78056	-1.10174	0.616994	-0.60187	-1.07334	1.71
13	1	3	5	11	4	35.26	-1.35114	-0.78056	-1.10174	0.616994	1.502517	1.737137	5
14	2	3	5	11	2	15.42	0.740115	-0.78056	-1.10174	0.616994	-0.60187	-0.49415	1.57
15	2	3	5	11	4	18.43	0.740115	-0.78056	-1.10174	0.616994	1.502517	-0.15564	3
16	1	3	5	11	2	14.83	-1.35114	-0.78056	-1.10174	0.616994	-0.60187	-0.56051	3.02
17	2	3	5	11	2	21.58	0.740115	-0.78056	-1.10174	0.616994	-0.60187	0.198627	3.92
18	1	3	5	11	3	10.33	-1.35114	-0.78056	-1.10174	0.616994	0.450322	-1.0666	1.67
19	2	3	5	11	3	16.29	0.740115	-0.78056	-1.10174	0.616994	0.450322	-0.39631	3.71
20	1	3	5	11	3	16.97	-1.35114	-0.78056	-1.10174	0.616994	0.450322	-0.31983	3.5
21	2	3	6	11	3	20.65	0.740115	-0.78056	-0.23251	0.616994	0.450322	0.094035	3.35

Multiple Linear Regression Model 2:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.684226
R Square	0.468166
Adjusted R Square	0.454645
Standard Error	1.022799
Observations	243

ANOVA

	df	SS	MS	F	ignificance F
Regression	6	217.3281	36.22135	34.62456	6.99E-30
Residual	236	246.8837	1.046117		
Total	242	464.2118			

	Coefficients	tandard Errc	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Jpper 95.0%
Intercept	3.002387	0.065613	45.75932	2.5E-119	2.873126	3.131648	2.873126	3.131648
sex (Standardized)	-0.017517	0.067734	-0.258616	0.796157	-0.150957	0.115923	-0.150957	0.115923
smoker (Standardize	-0.035332	0.068415	-0.516438	0.606032	-0.170115	0.09945	-0.170115	0.09945
day (Standardized)	-0.060979	0.138723	-0.439577	0.660646	-0.334273	0.212314	-0.334273	0.212314
time (Standardized)	-0.05184	0.137885	-0.375968	0.707278	-0.323483	0.219803	-0.323483	0.219803
size (Standardized)	0.166285	0.084936	1.957776	**0.051435**	-0.001044	0.333614	-0.001044	0.333614
total_bill (Standardiz	0.838165	0.085017	9.858787	**2.01E-19**	0.670675	1.005654	0.670675	1.005654

SUI

Mu

R S

Adj

Sta

Obs

AN

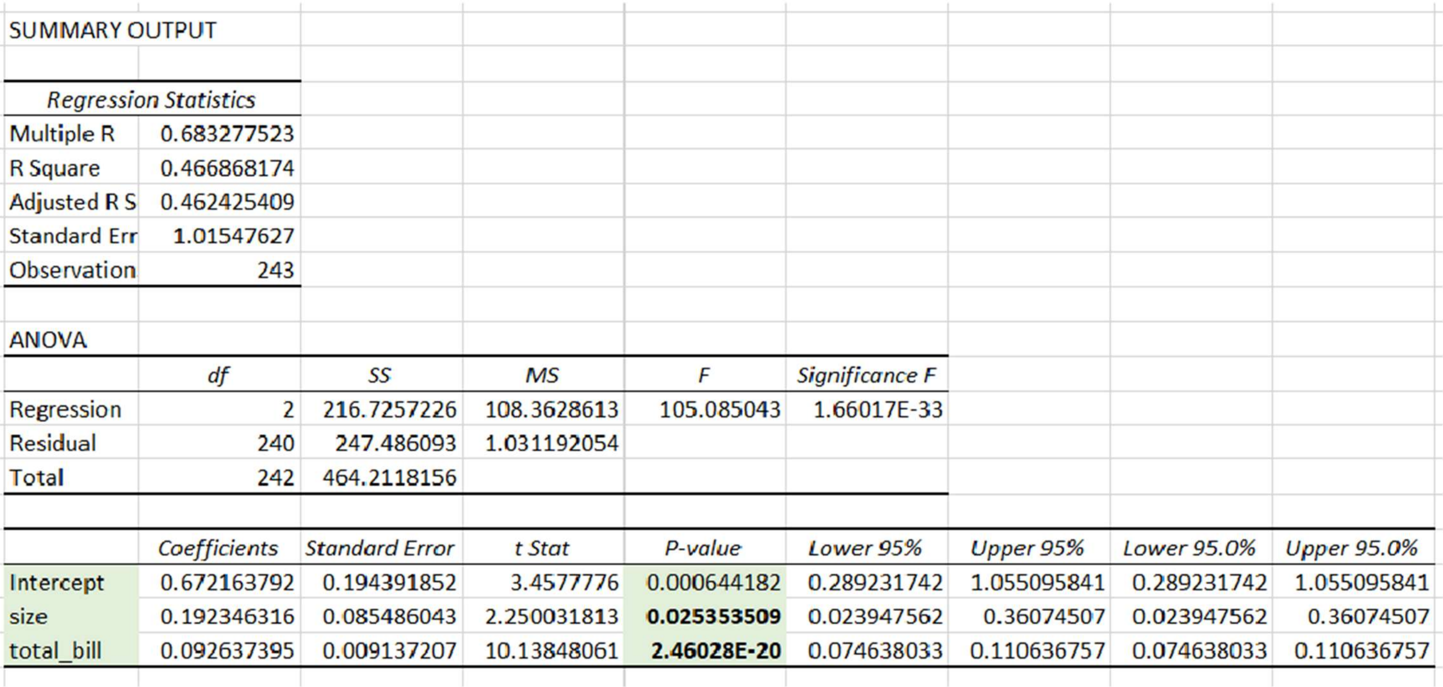
Reg

Res

Tot

	ower 95.0%	Upper 95.0%
Inte	-5.2579809	10.39269453
sex	-0.31568889	0.242423864
smo	-0.35072488	0.205036218
day	-0.2905611	0.184550488
tim	-0.72387672	0.491864201
size	-0.00109846	0.351027206
tota	).075426978	0.113100059

Multiple Linear Regression Model 3:



SUMMARY OUTPUT

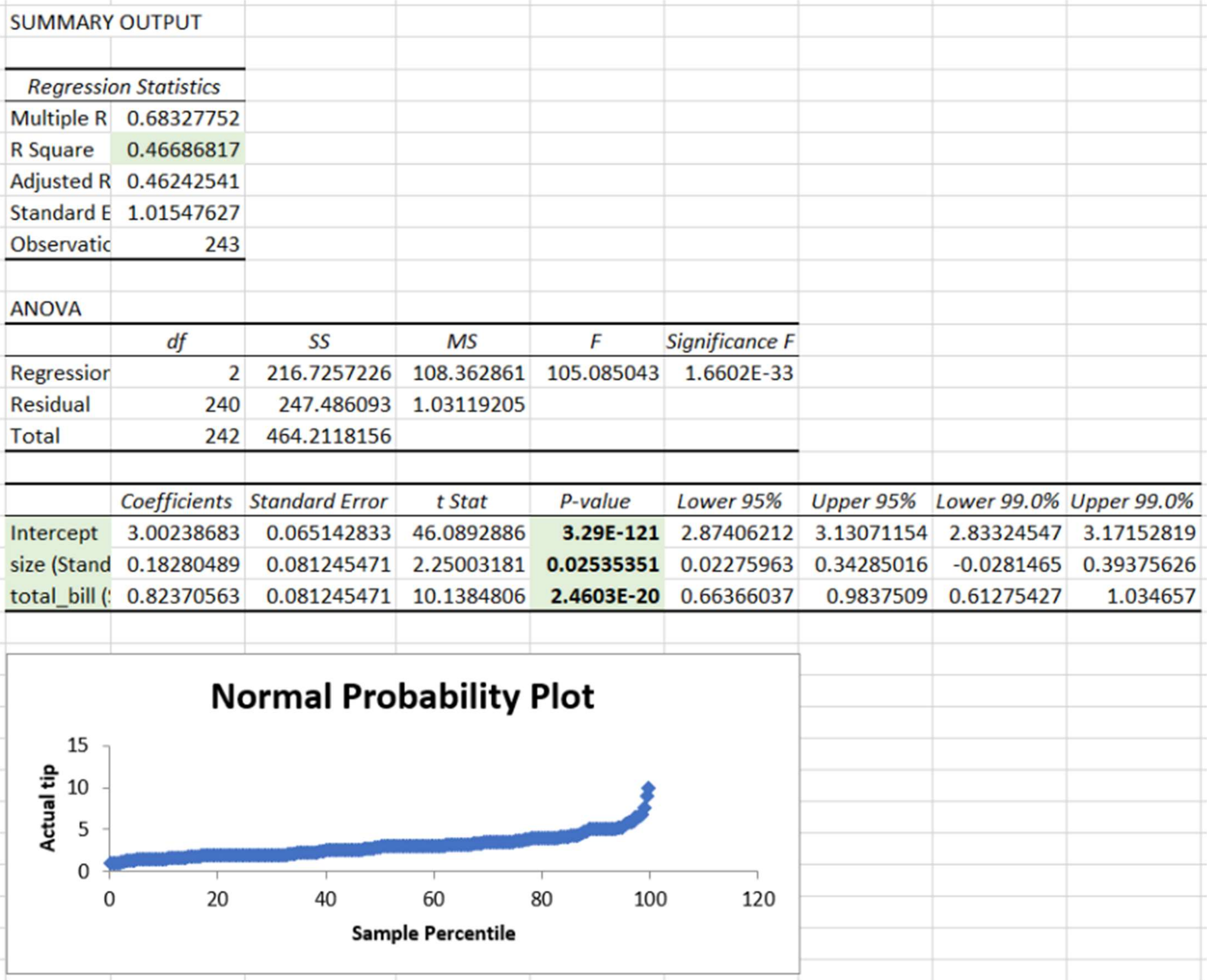
Regression Statistics	
Multiple R	0.683277523
R Square	0.466868174
Adjusted R S	0.462425409
Standard Err	1.01547627
Observation	243

ANOVA

	df	SS	MS	F	Significance F
Regression	2	216.7257226	108.3628613	105.085043	1.66017E-33
Residual	240	247.486093	1.031192054		
Total	242	464.2118156			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0.672163792	0.194391852	3.4577776	0.000644182	0.289231742	1.055095841	0.289231742	1.055095841
size	0.192346316	0.085486043	2.250031813	**0.025353509**	0.023947562	0.36074507	0.023947562	0.36074507
total_bill	0.092637395	0.009137207	10.13848061	**2.46028E-20**	0.074638033	0.110636757	0.074638033	0.110636757

Multiple Linear Regression Model 4:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.68327752
R Square	0.46686817
Adjusted R	0.46242541
Standard E	1.01547627
Observatic	243

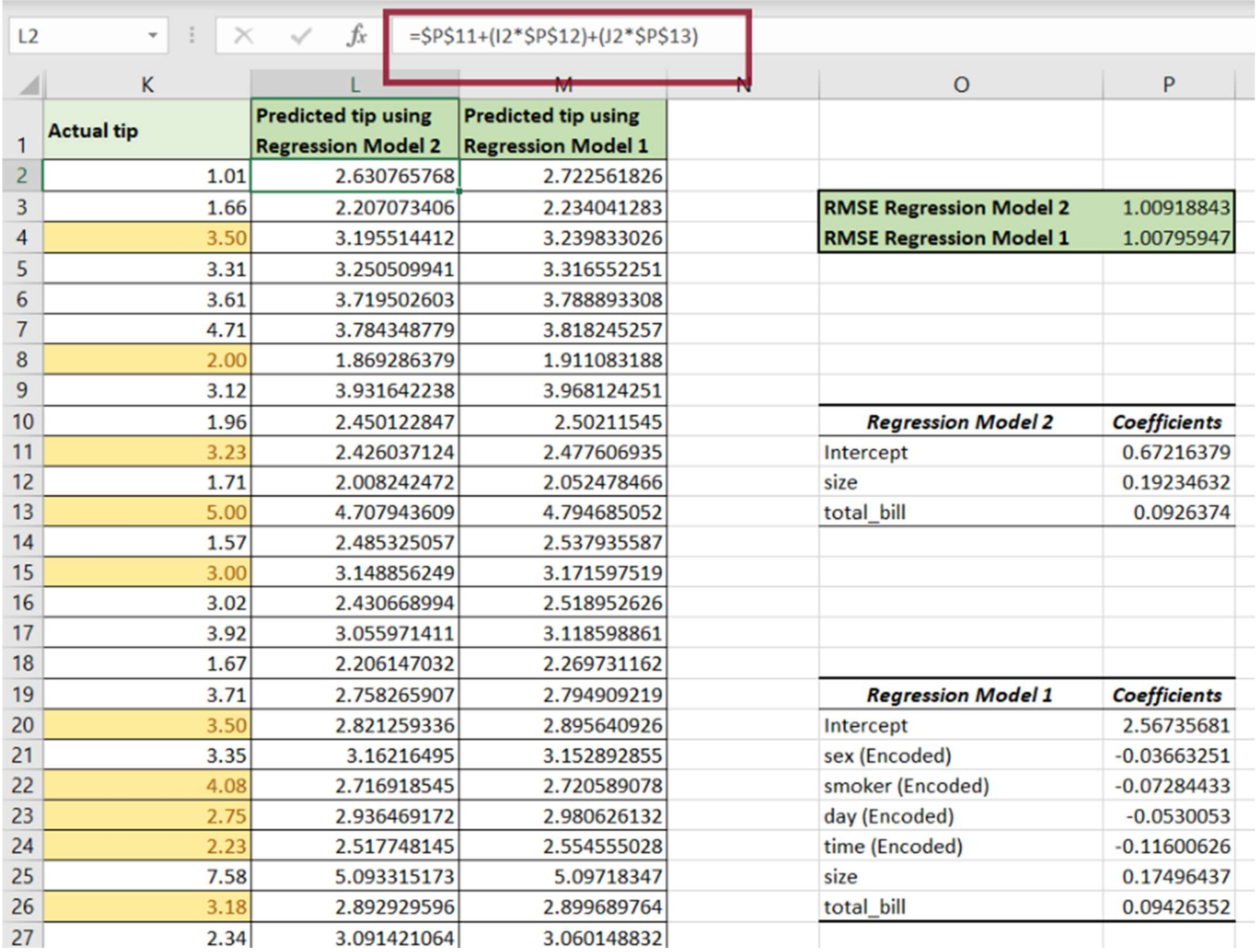
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regressior	2	216.7257226	108.362861	105.085043	1.6602E-33
Residual	240	247.486093	1.03119205		
Total	242	464.2118156			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	3.00238683	0.065142833	46.0892886	**3.29E-121**	2.87406212	3.13071154	2.83324547	3.17152819
size (Stand	0.18280489	0.081245471	2.25003181	**0.02535351**	0.02275963	0.34285016	-0.0281465	0.39375626
total_bill (	0.82370563	0.081245471	10.1384806	**2.4603E-20**	0.66366037	0.9837509	0.61275427	1.034657

Comparison of Predicted Tip values using the models above with the Actual values using RMSE:

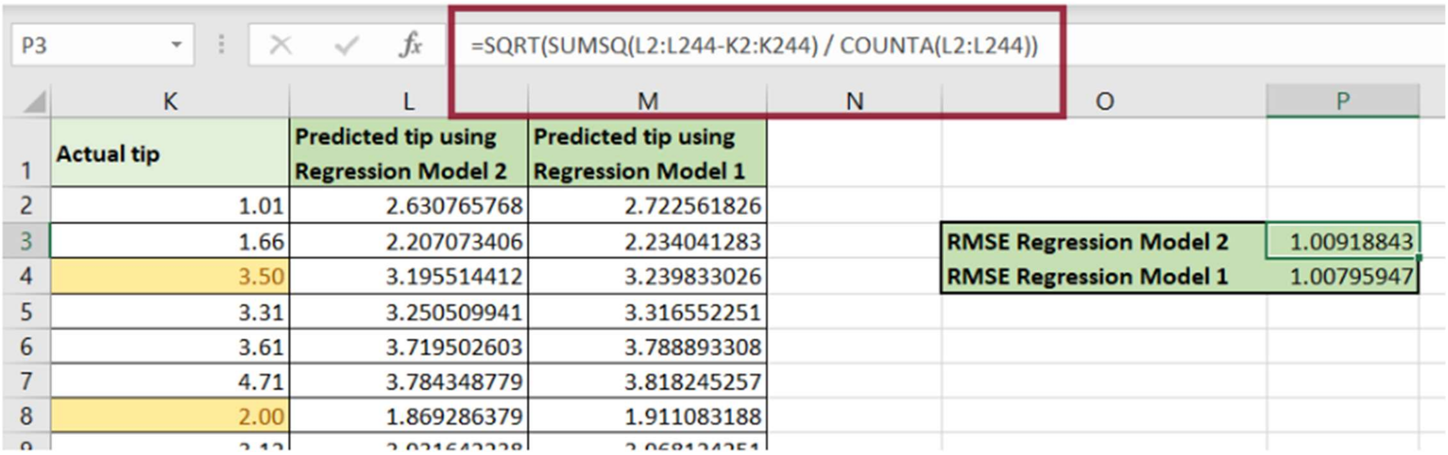
Prediction using Model 1 and 2:



L2 =P11+(I2*P12)+(J2*P13)

	K	L	M	N	O	P
1	Actual tip	Predicted tip using Regression Model 2	Predicted tip using Regression Model 1			
2	1.01	2.630765768	2.722561826			
3	1.66	2.207073406	2.234041283		RMSE Regression Model 2	1.00918843
4	3.50	3.195514412	3.239833026		RMSE Regression Model 1	1.00795947
5	3.31	3.250509941	3.316552251			
6	3.61	3.719502603	3.788893308			
7	4.71	3.784348779	3.818245257			
8	2.00	1.869286379	1.911083188			
9	3.12	3.931642238	3.968124251			
10	1.96	2.450122847	2.50211545		*Regression Model 2*	*Coefficients*
11	3.23	2.426037124	2.477606935		Intercept	0.67216379
12	1.71	2.008242472	2.052478466		size	0.19234632
13	5.00	4.707943609	4.794685052		total_bill	0.0926374
14	1.57	2.485325057	2.537935587			
15	3.00	3.148856249	3.171597519			
16	3.02	2.430668994	2.518952626			
17	3.92	3.055971411	3.118598861			
18	1.67	2.206147032	2.269731162			
19	3.71	2.758265907	2.794909219		*Regression Model 1*	*Coefficients*
20	3.50	2.821259336	2.895640926		Intercept	2.56735681
21	3.35	3.16216495	3.152892855		sex (Encoded)	-0.03663251
22	4.08	2.716918545	2.720589078		smoker (Encoded)	-0.07284433
23	2.75	2.936469172	2.980626132		day (Encoded)	-0.0530053
24	2.23	2.517748145	2.554555028		time (Encoded)	-0.11600626
25	7.58	5.093315173	5.09718347		size	0.17496437
26	3.18	2.892929596	2.899689764		total_bill	0.09426352
27	2.34	3.091421064	3.060148832			

Comparison using RMSE (Root Mean Square Error):



P3 =SQRT(SUMSQ(L2:L244-K2:K244) / COUNTA(L2:L244))

	K	L	M	N	O	P
1	Actual tip	Predicted tip using Regression Model 2	Predicted tip using Regression Model 1			
2	1.01	2.630765768	2.722561826			
3	1.66	2.207073406	2.234041283		RMSE Regression Model 2	1.00918843
4	3.50	3.195514412	3.239833026		RMSE Regression Model 1	1.00795947
5	3.31	3.250509941	3.316552251			
6	3.61	3.719502603	3.788893308			
7	4.71	3.784348779	3.818245257			
8	2.00	1.869286379	1.911083188			

RESULT AND CONCLUSION: Regression Model 2 has better RMSE and lesser R-squared error than Regression Model