Sephora Case Study Questions 2

1. Are these media channels data correlated? List any relationships you find concerning.

Correlation Heat Map

Correlation Table Between the Media Campaigns

	SALES	*Social Impres*	*Search Impres*	*Affiliate Clicks*
SALES	1			
Paid Social Impressions	0.54948877	1		
Paid Search Impressions	0.75367823	0.56997532	1	
Affiliate Clicks	0.87867948	0.55372593	0.66468822	1

Yes, as displayed on the correlation heat map there are correlations between the performance of several media channels. There exists positive correlation between Sales and Paid Search Impressions (.75), Paid Social Impressions (.55), and Affiliate Clicks (.88). Between the social media campaigns, there exists a positive correlation between Paid Social Impressions and Paid Search Impressions (.57), Paid Social Impressions and Affliate

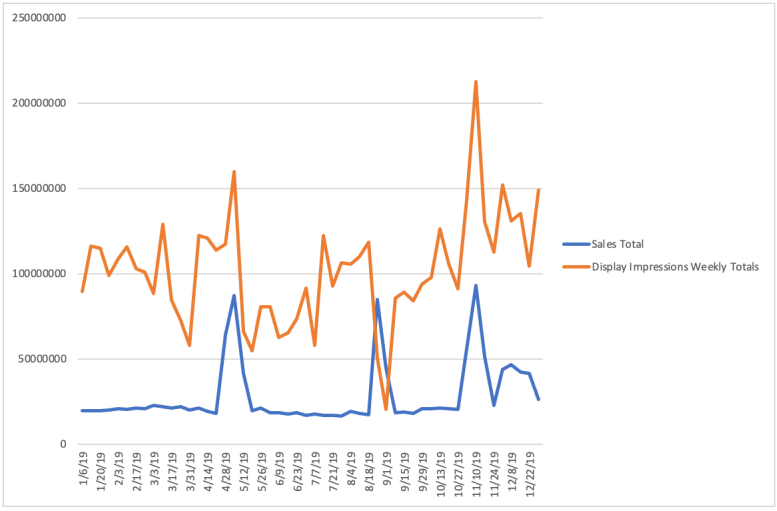
Cicks (.55) and Affiliate Clicks and Paid Search Impressions (.66). One area of concern could be the strong correlation between Affiliate Clicks and 'SALES' (.88) as it could suggest an indication of a collinearity problem.

2. How much sales was driven from each channel in each week? What was the ROAS of each channel?

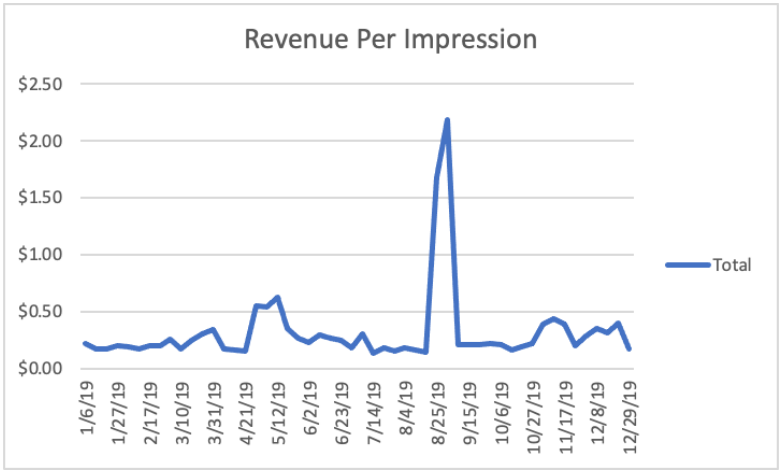
To calculate sales driven by channel per week, I found each channel's revenue per impression by dividing the total sales by each media campaign's corresponding impression count. Then, to calculate the ROAS of each channel, I used the ROAS formula and divided the sales by each media campaign's corresponding spend.

Display Impressions Summary

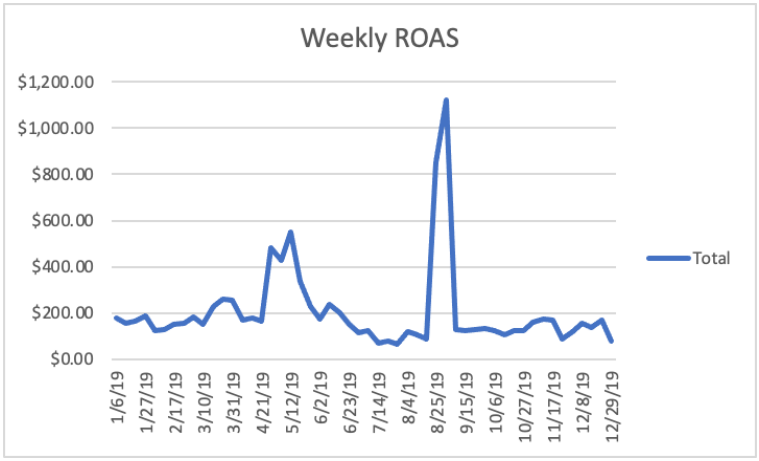
Total Weekly Sales to Display Impressions



Weekly Revenue Per Impression (Display Impressions)

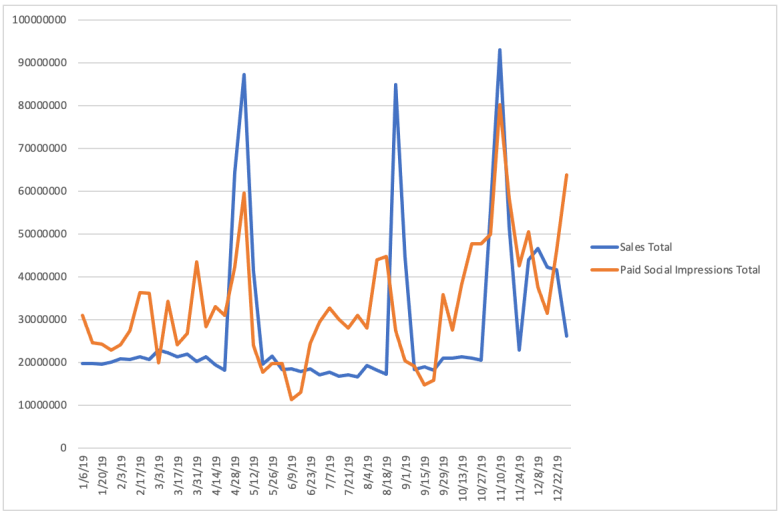


Weekly ROAS (Display Impressions)

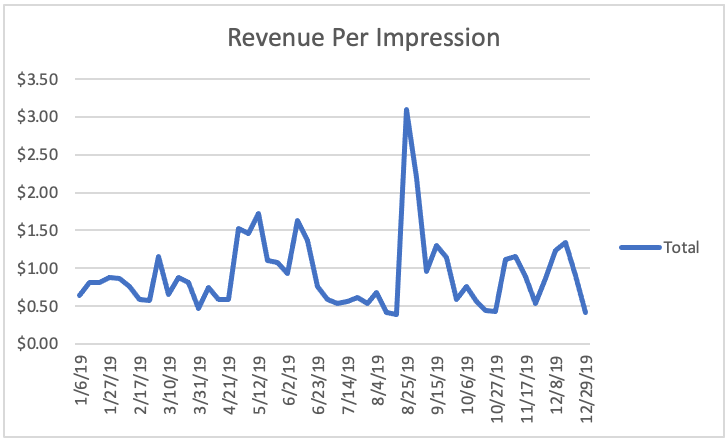


Paid Social Impressions Summary

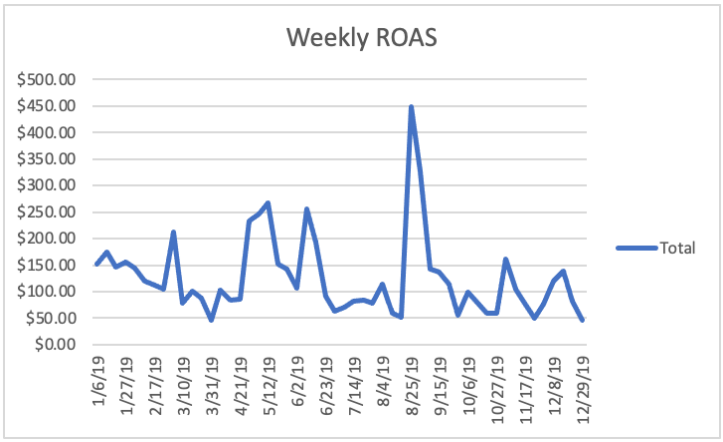
Total Weekly Sales to Paid Social Impressions



Weekly Revenue Per Impression (Paid Social Impressions)

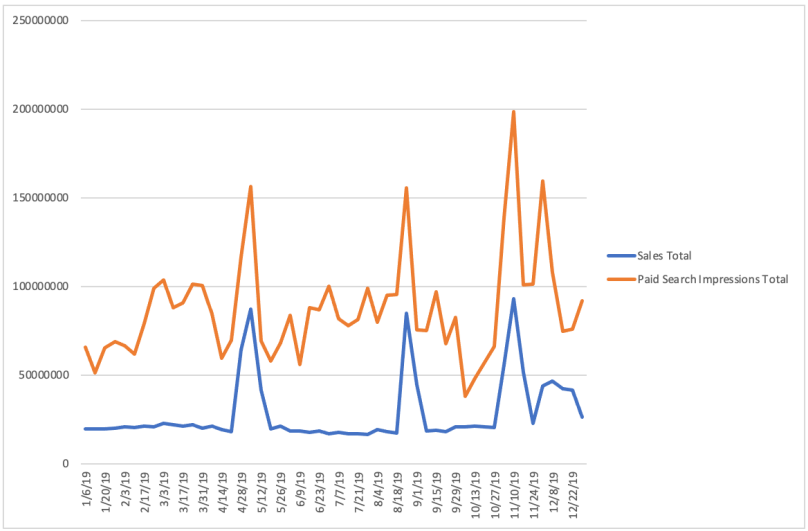


Weekly ROAS (Paid Social Impressions)

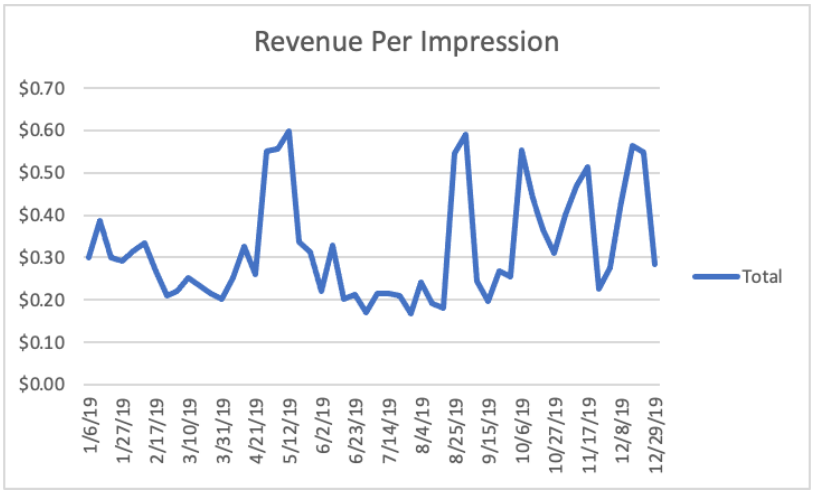


Paid Search Impressions Summary

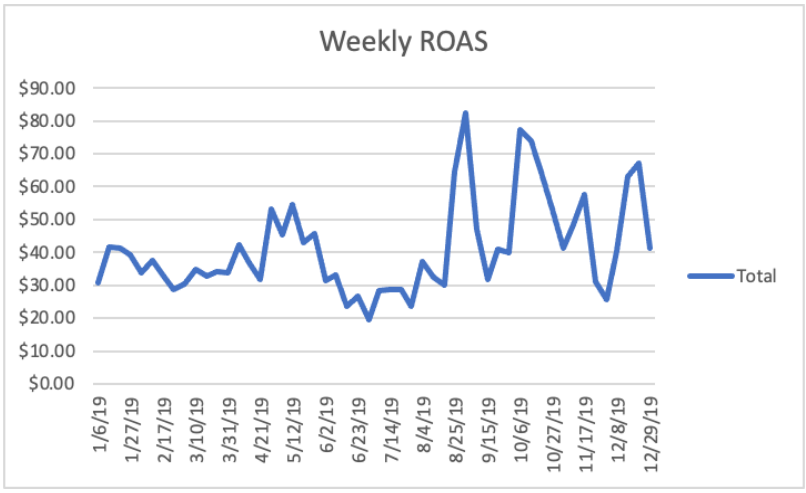
Total Weekly Sales to Paid Search Impressions



Weekly Revenue Per Impression (Paid Search Impressions)

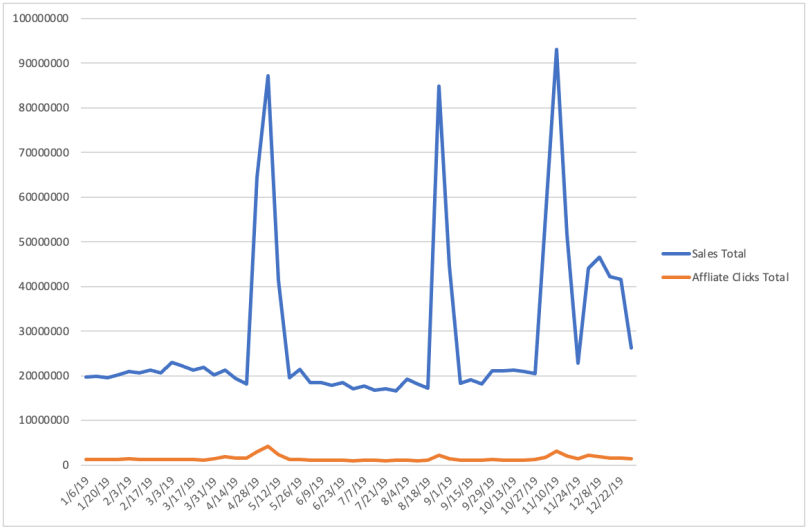


Weekly ROAS (Paid Search Impressions)

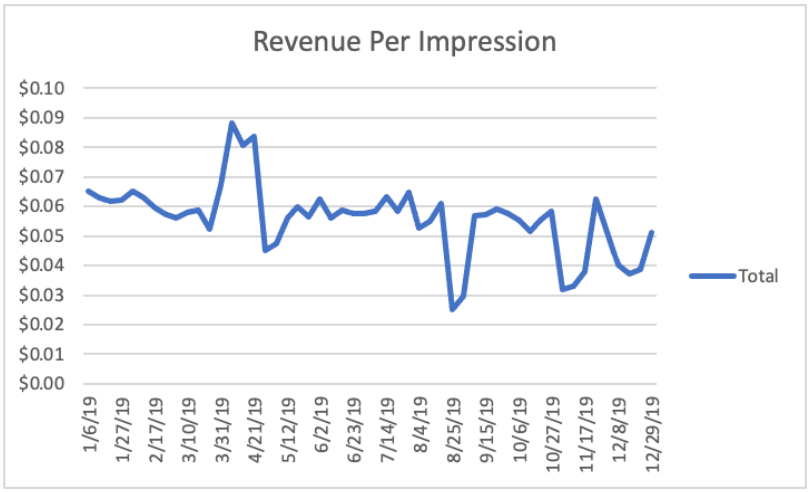


Affiliate Clicks Summary

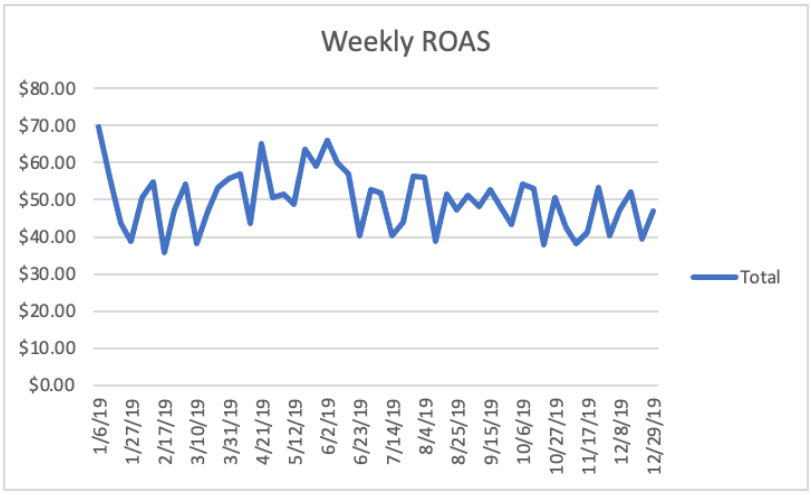
Total Weekly Sales to Affiliate Clicks



Weekly Revenue Per Impression (Affiliate Clicks)



Weekly ROAS (Affiliate Clicks)



Multiple Linear Regression Model (Python)

```
:
import seaborn as sb
import matplotlib.pyplot as plt
from matplotlib import style
import statsmodels.api as sm
X = ddData[["Paid Social Impressions", "Affiliate Clicks", "Display Impressions", "Paid Search Impressions"]]
y = ddData['SALES']
sm_X1_var = sm.add_constant(X)

mlr_model = sm.OLS(y, sm_X1_var)
mlr_reg = mlr_model.fit()
print(mlr_reg.summary())
```

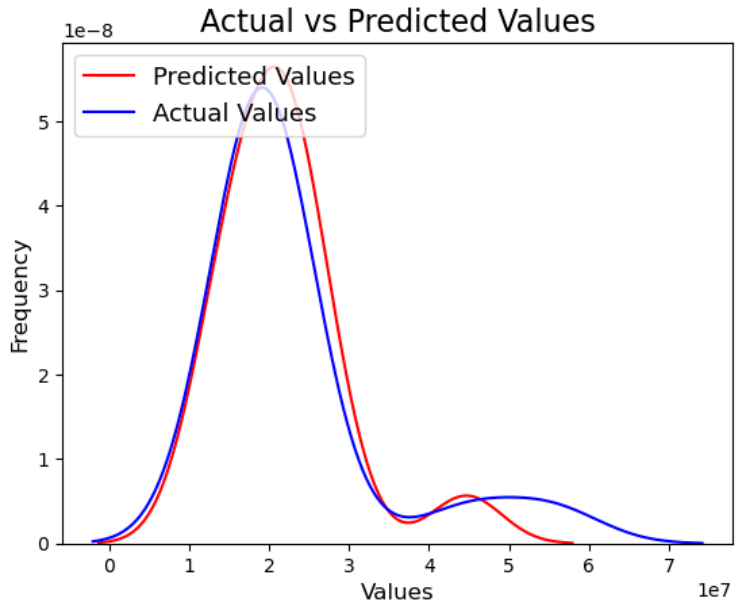
OLS Regression Results

Dep. Variable:	SALES	R-squared:	0.836
Model:	OLS	Adj. R-squared:	0.822
Method:	Least Squares	F-statistic:	60.08
Date:	Fri, 01 Dec 2023	Prob (F-statistic):	6.89e-18
Time:	09:06:50	Log-Likelihood:	-896.91
No. Observations:	52	AIC:	1804.
Df Residuals:	47	BIC:	1814.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.42e+07	4.17e+06	-3.402	0.001	-2.26e+07	-5.8e+06
Paid Social Impressions	0.1507	0.127	1.185	0.242	-0.105	0.406
Affiliate Clicks	22.0236	2.617	8.415	0.000	16.759	27.289
Display Impressions	-0.0965	0.050	-1.916	0.061	-0.198	0.005
Paid Search Impressions	0.1819	0.052	3.515	0.001	0.078	0.286

Omnibus:	8.663	Durbin-Watson:	1.119
Prob(Omnibus):	0.013	Jarque-Bera (JB):	8.175
Skew:	0.763	Prob(JB):	0.0168
Kurtosis:	4.202	Cond. No.	5.50e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.5e+08. This might indicate that there are
strong multicollinearity or other numerical problems.



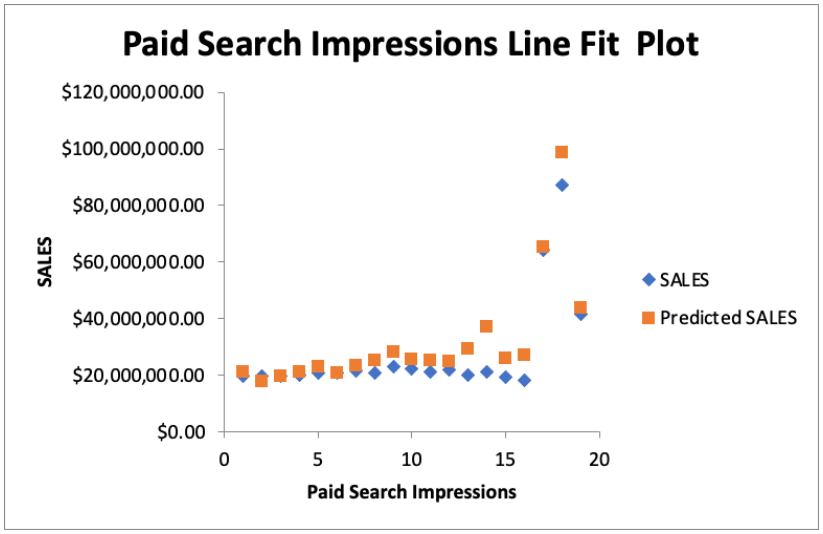
Multiple Linear Regression Model (Excel)

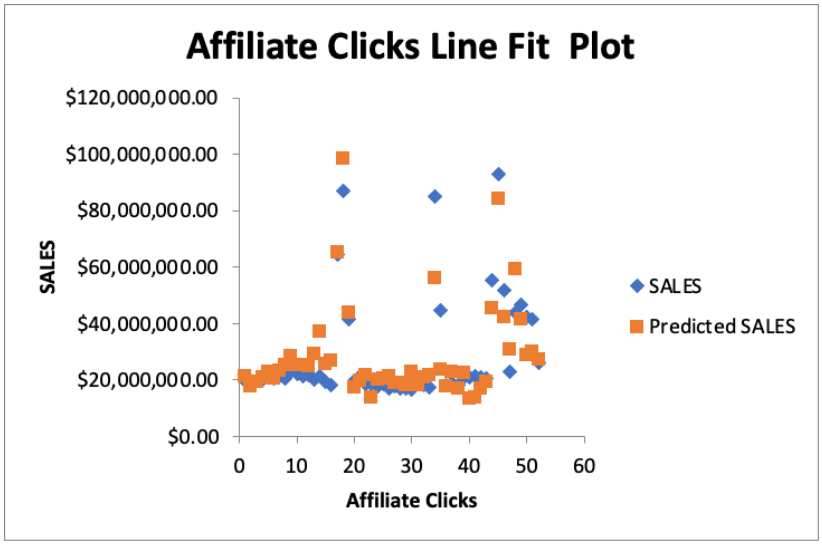
Regression Statistics	
Multiple R	0.90754132
R Square	0.82363125
Adjusted R Square	0.81260821
Standard Error	8096189.07
Observations	52

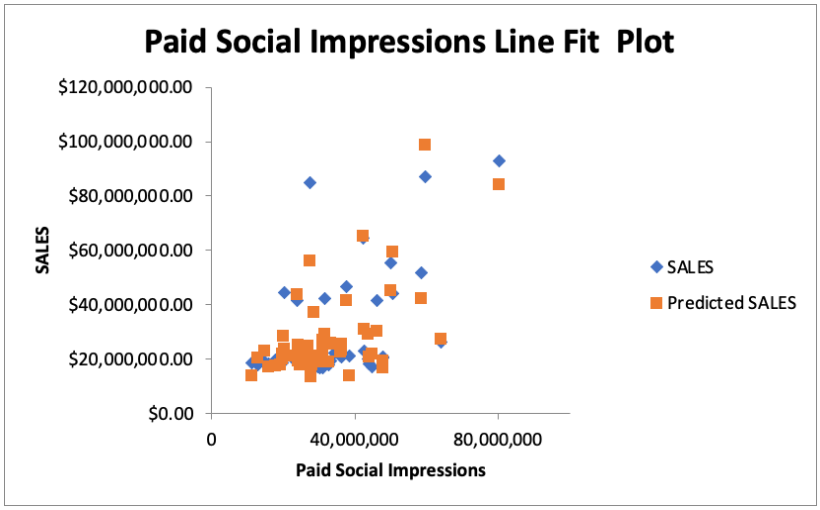
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1.4693E+16	4.8977E+15	74.7190207	4.1984E-18
Residual	48	3.1463E+15	6.5548E+13		
Total	51	1.7839E+16			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-18388560.12	3654040.071	-5.0323915	7.2189E-06	-25735500.1	-11041620	-25735500	-11041620
Paid Social Impressions	0.003440857	0.104047053	0.0330702	0.97375574	-0.205759765	0.21264148	-0.2057598	0.21264148
Paid Search Impressions	0.187453171	0.053071282	3.53210182	0.00092235	0.080746207	0.29416013	0.08074621	0.29416013
Affiliate Clicks	21.12528094	2.645476757	7.98543434	2.3166E-10	15.80619343	26.4443685	15.8061934	26.4443685







3. Do you have an intercept for your model? How would you explain it?
Yes there is a model intercept in my Python model, represented by the term 'const' and it has a value of -1.42e+07, I added this intercept from the Statsmodel library in Python, using the add_constant() function to add a constant in order to allow for a more complete model. This intercept is the estimated value of the dependent variable, in this case, 'SALES' when all the independent variables (the campaigns) are nonexistent. This constant shows the sales that occur when no media campaigns occur. In this case, this negative coefficient suggests that the intercept is nonsensical as it is extremely unlikely that all the media campaigns would have a customer impression of zero.

The Excel linear regression model has an negative intercept as well, demonstrating a nonsensical meaning because, as mentioned previously, it seems quite unlikely that all the media campaigns would have no impression on any customers.

4. How good does your model fit?
The Python linear regression model fits relatively well, as it has an r-squared value of 0.836 signifying that 83.6% of the variability exhibited in total sales can be attributed to the media campaigns. However, between the various media campaigns, only “Affiliate Clicks” has a statistically positive impact on sales with a $P< 0.001$. Furthermore, according to the distribution plot I created that maps the actual values in comparison to the predicted values calculated using the MLR model, the prediction values have performed generally precisely to the actual values.

The Excel linear regression model fits relatively well, with an adjusted r-square value of 0.81 or 81%, demonstrating that 81% of the independent variable, ‘SALES’, is explained by the dependent variables, the media campaigns. Moreover, in terms of the media campaigns and their correlations with total sales, the p-value of Paid Social Impressions is .97 indicating that it is not statistically significant. The Paid Search Impressions has a p-value of 0.0009 indicating that it is less than standard significance level of 0.05 and is significant to predicting sales. Lastly, Affiliate Clicks has a very small p-value, indicating high statistical significance to predicting sales. Furthermore, the line of best fit graphs breakdown the predicted to actual sales by media campaign and reinforce the accuracy of the model as the predicted sales for each media campaign aligns similarly to the actual sales.

5. Is there any other information to consider to improve the model?
There is a high condition number in the Python linear regression model, which suggests that there may be an issue with multicollinearity and possible redundant predictors may need to be removed to improve the model. In order to improve the model and the prediction on the correlation between sales and each media campaign, perhaps more information could be introduced. For instance, time of day may be a relevant predictor on user engagement with media campaigns as consumers may be more likely to buy goods at a certain time, perhaps outside of general working hours. Another factor to consider would be geometric location, as certain social media campaigns may resonate more or less with specific consumers based on their social community environment. Furthermore, to give more detail to the resulting sales from each social media campaign, the factor of number of orders could be considered in order to reveal if certain social media campaigns resulted in more or less orders.