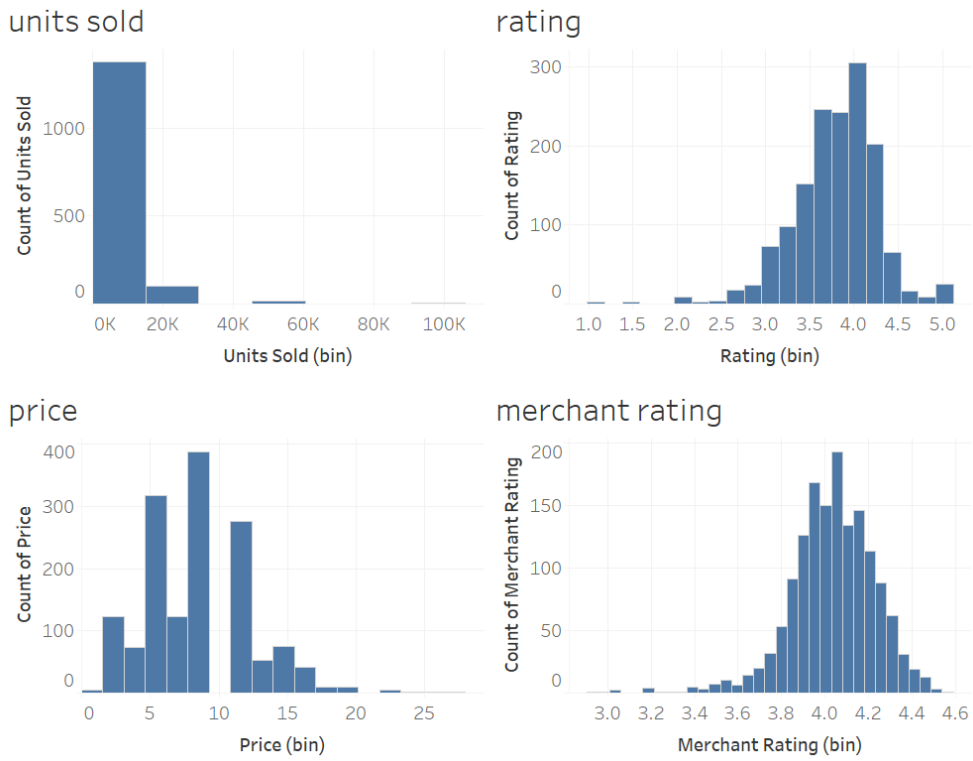
Harshaanth Thiyagaraja Kumar

Linear Regression -Predicting product sales

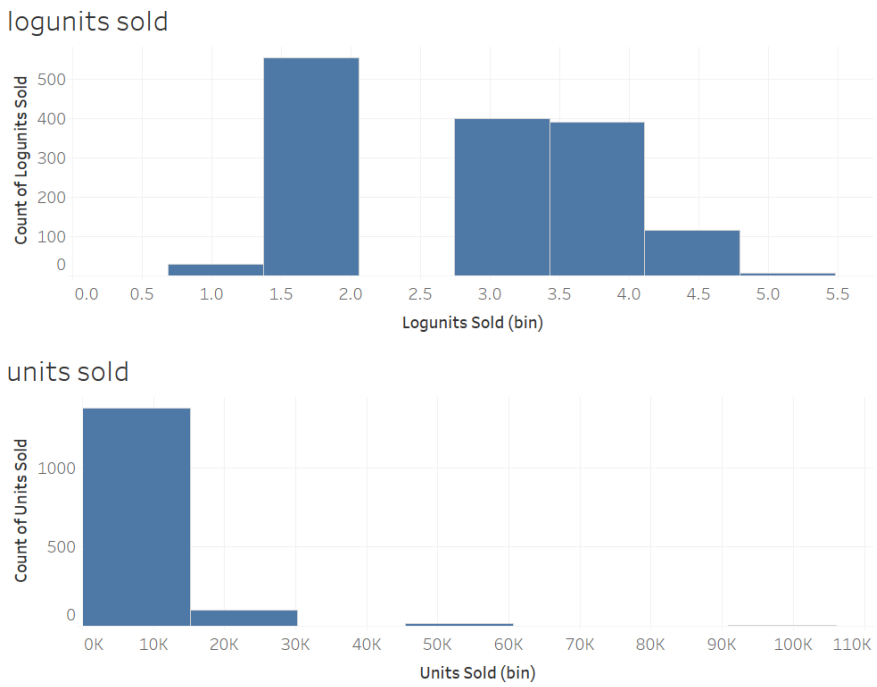
Use the file of summer products with rating and performance to build a model and find factors that could affect product sales.
Variable definitions could be found in the file.

1. Show histograms of all numerical variables. Considering we will run a linear regression, should we do log transformations of any variables? Why? (paste the screenshots)



I will do a log transformation of units sold, because it's highly skewed. And others look like a normal distribution.

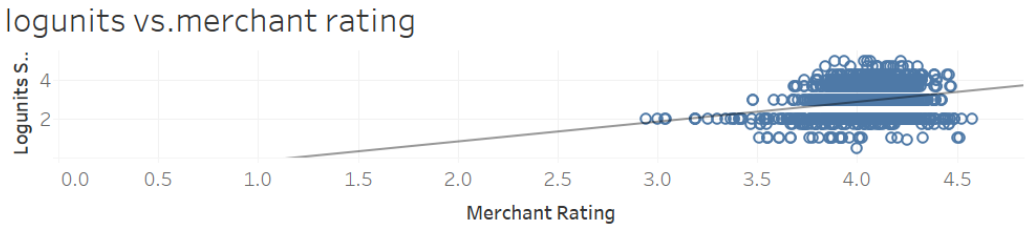
2. Do a log transformation of units_sold (use log10(units_sold+1)). Show the histogram of logunits_sold. Comparing to the histogram of units_sold, any observations? (paste the screenshots)

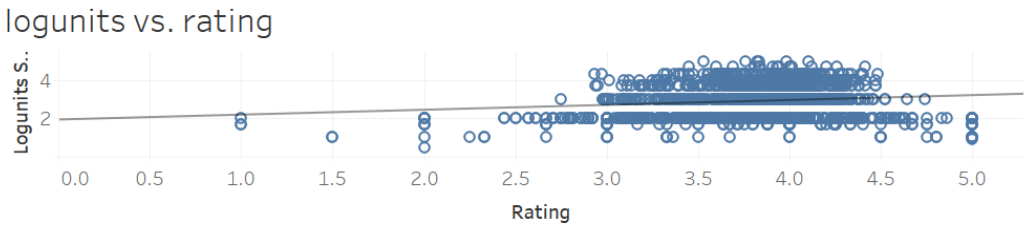


After log transformation, the distribution is more like a normal distribution.

3. To make it simple and clean, delete units_sold
4. Create the scatter plots of logunits_sold and all the other numerical variables and show the trend line, any trend? If yes, please explain the trend. (paste screenshots)

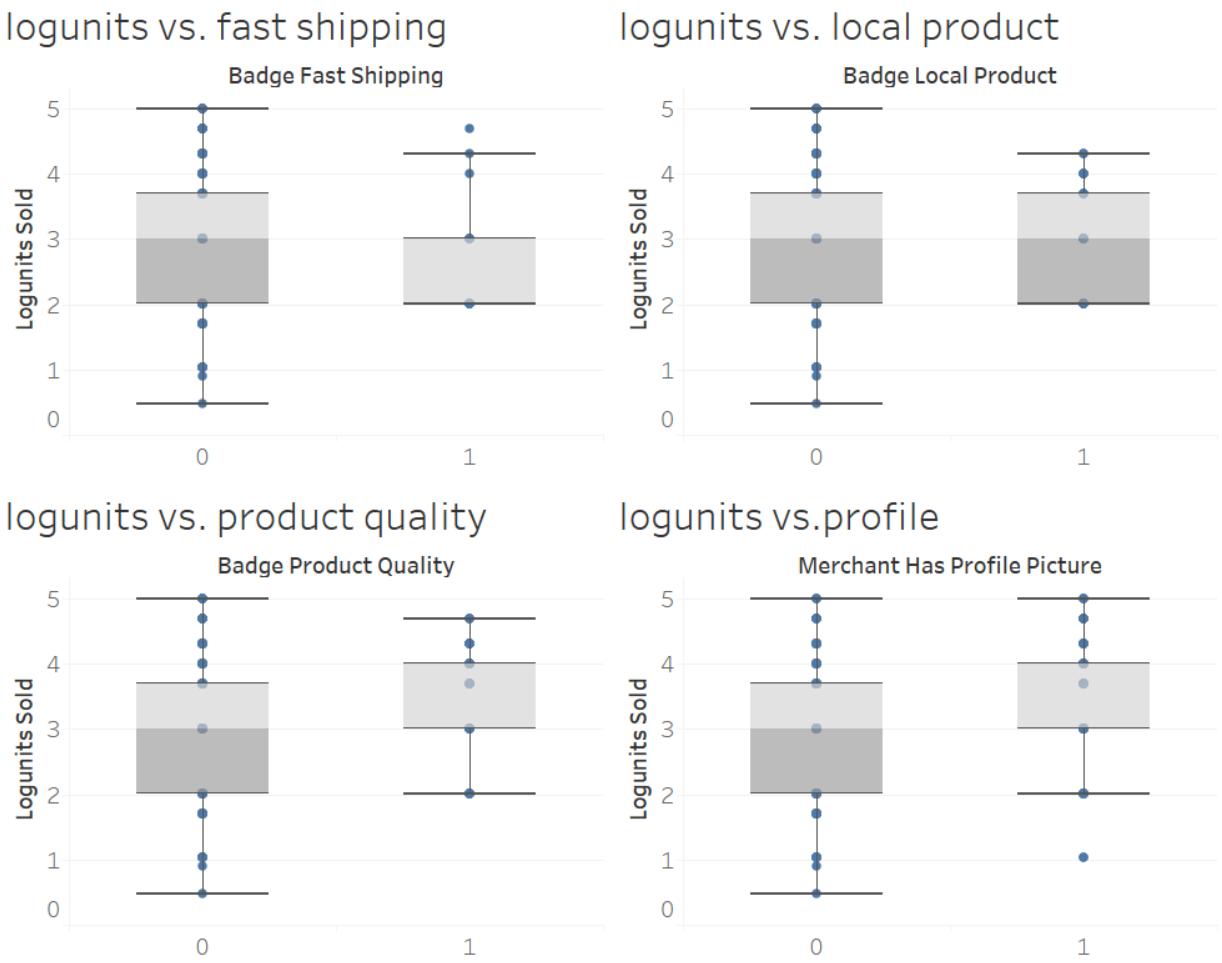
The three variables are all positively correlated with logunits_sold.





5. Create the box plots of logunits_sold and all the other categorical variables, any observations? If yes, please explain. (paste screenshots)

If we compare the median, it seems that the badges of local product, product quality don't make any different. It doesn't matter whether the merchant has a profile picture. And it seems that the median is higher with the badge of fast shipping.



6. Run the linear regression using logunits_sold as the dependent variable, and other variables as independent variables. Split data into training and validation (60% and 40%), and use stepwise for variable selection.
    a. Report the results using all predictors. Report this model's
        (1) *Regression equation* (write down the model and paste the table on Linreg_output sheet),
        Logunits_sold=-1.49-0.02badeg_localproduct+0.14badge_productquality-0.27badge_fastshipping+0.36mechant_profile_picture+0.008price+0.18rating+0.89mechant_rating

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	-1.49601228	-2.743678068	-0.248346486	0.63571115	-2.35328935	0.0188246
badge_local	-0.02075127	-0.537906997	0.49640446	0.263501384	-0.07875203	0.9372475
badge_produ	0.14357843	-0.087413794	0.374570657	0.117695247	1.21991699	0.2228191
badge_fast_s	-0.27227993	-0.879918667	0.335358799	0.309604319	-0.87944488	0.3793972
merchant_ha	0.36220814	0.19876189	0.525654382	0.083279194	4.349323282	1.523E-05
price	0.0080759	-0.006937644	0.023089446	0.007649707	1.055713735	0.2913852
rating	0.17519844	0.051644491	0.298752399	0.062953258	2.782992516	0.0054999
merchant_ra	0.89221114	0.575269338	1.209152934	0.161488306	5.524927217	4.328E-08

(2) Which variables are significant?

Merchant_profile_picture, rating, merchant_rating

(3) Compare the model performances of Training and Validation data

training

Metric	Value
SSE	694.6522701
MSE	0.772694405
RMSE	0.879030378
MAD	0.754325564
R2	0.088876388

validation

Metric	Value
SSE	441.122529
MSE	0.735204215
RMSE	0.857440502
MAD	0.733891859
R2	0.050581185

RMSE of the validation is a little bit smaller compared to RMSE of the training data, but the r2 is lower. It seems like the model has good predictive power on the validation data, but worse explanatory power. In general, the model would work well on new dataset for prediction.

b. Choose the best model among models recommended by variable selection. Report the selected model's

(1) Regression Model (write down the model and paste the table on Linreg_output sheet),

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	-1.55888367	-2.781290045	-0.336477292	0.62284475	-2.50284468	0.0124972
merchant_has_profile_picture	0.35676615	0.193911745	0.519620552	0.082978142	4.299519598	1.9E-05
rating	0.19916685	0.080249688	0.318084014	0.060591087	3.287065186	0.0010518
merchant_rating	0.90415183	0.589908911	1.218394751	0.160114146	5.646920357	2.193E-08

Logunits_sold=-1.56+0.36mechant_profile_picture+0.2rating+0.90mechant_rating

(2) Compared to the model reported in a, which variables are removed?

badeg_localproduct, badge_productquality, badge_fastshipping, price

(3) Compare the model performances of Training and Validation data

Training: Prediction Summary

Metric	Value
SSE	697.2521198
MSE	0.77558634
RMSE	0.880673799
MAD	0.754593941
R2	0.085466358

Validation: Prediction Summary

Metric	Value
SSE	443.9514507
MSE	0.739919084
RMSE	0.860185494
MAD	0.737126317
R2	0.044492556

The comparison is pretty similar as what we did in part a.

RMSE of the validation is a little bit smaller compared to RMSE of the training data, but the r2 is lower. It seems like the model has good predictive power on the validation data, but worse explanatory power. In general, the model would work well on new dataset for prediction.

c. Compare the predictive accuracy of both models (a) and (b) using measures such as RMSE. Based on the outcomes, which model would you choose, and why? (hints: considering the reasons why we do variable selection)

I will choose model b.

The performance of a and b are pretty similar. However, b is simpler. There are no insignificant variables. And there is no potential issue of multi-collinearity.

d. If a product is $2, with a local product badge, no quality badge, no fast shipping badge, rating of 4.5. The merchant doesn't have a profile picture and the average merchant rating is 4.8. What's the number of predicted units_sold?

Logunits_sold=-1.56+0.36mechant_profile_picture+0.2rating+0.90mechant_rating

*=-1.56+0.36*0 +0.2*4.5+0.90*4.8*

=3.66 (this number could be a little bit different if they use different number of decimals)

Units_solde=$10^{3.66}$-1=4569