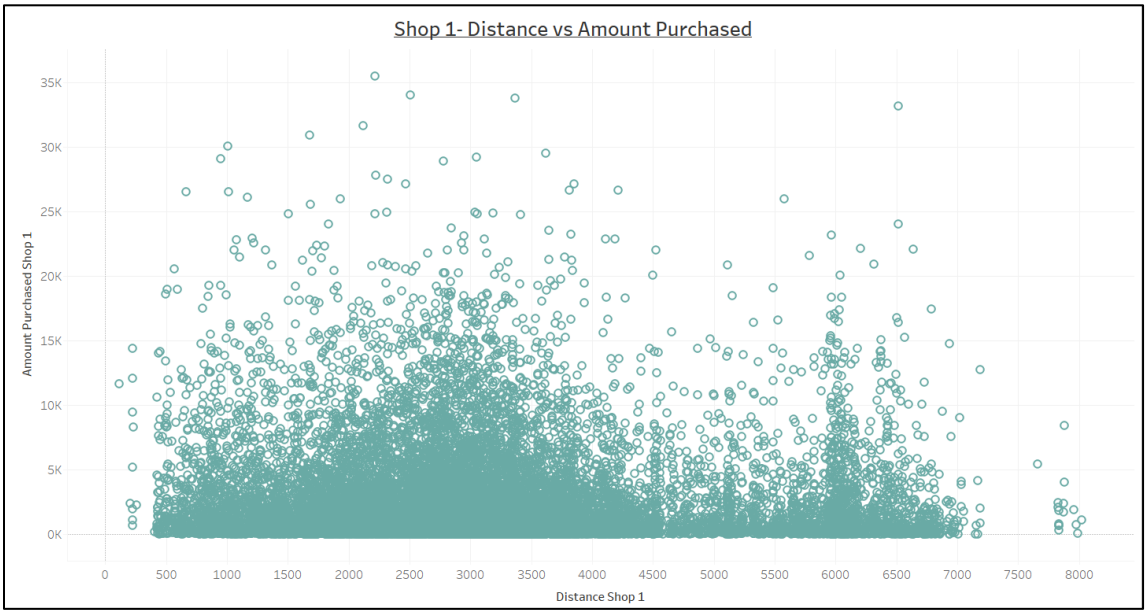
Homework #4
MEM 410, Managerial Analytics, Winter 2018
Due: Start of Class 2/22/18
Electronic Submission Only

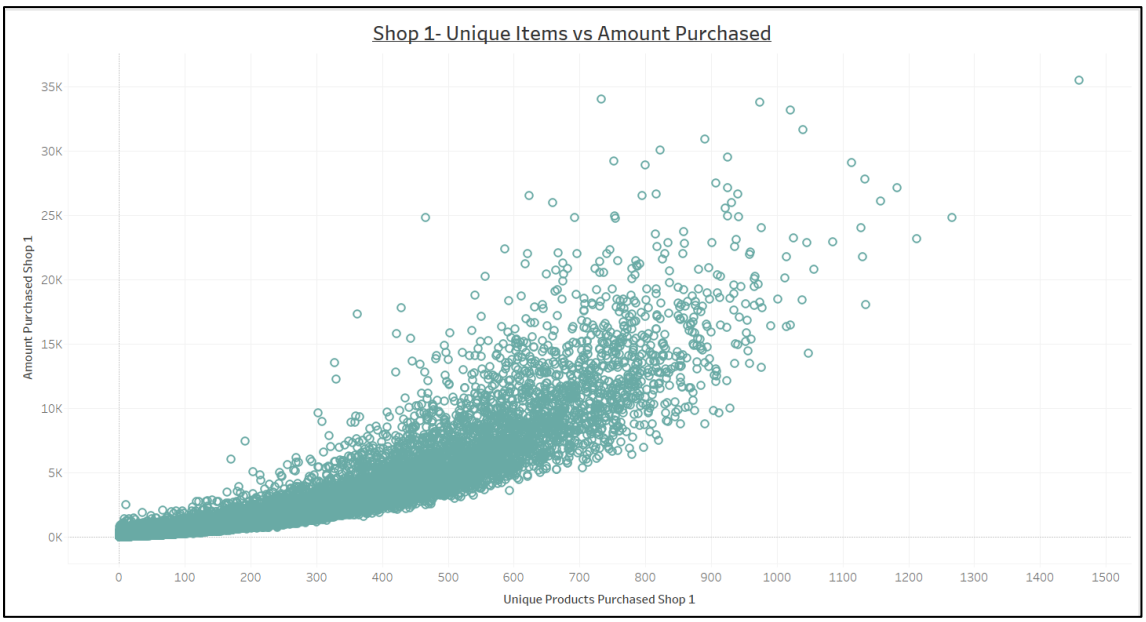
Download the "Retail Data" from Canvas. This dataset contains aggregated customer information in an area with five different shops. Various data attributes including distance to the shop, the amount of purchase, and price etc. are included in the dataset. The headings are rather self-explanatory. We will be using this dataset for the entire homework.

1. [5 points] When it comes to model building, it usually takes significant amount of time to organize the data so that one can start modeling. In many cases, we will need to think how to transform raw data into variables that are more meaningful in a model. Let's assume for the moment that the owner of Shop 1 would like to better understand its competitive relationship with Shop 2. Use Access (or any other software that you are comfortable with), perform the following
   - Construct a dataset that consists of only the customers who shop in Shop 1 and in Shop 2 and nowhere else. Also in this dataset, include information about distance, price, total purchase amount, number of unique items purchased related to Shop 1 and Shop 2 respectively
   - As discussed in lecture, understanding univariate relationship is usually helpful prior to building any models. Construct scatter plots with amount purchased in shop1 as the y-axis, and with the following variable as the x-axis
     - Distance to Shop 1
     - Average Price at Shop 1
     - Number of unique items at Shop 1
   - Submit these univariate plots
   - Does any of these indicate strong correlation?



Since the scatterplot is pretty spread out, there is no strong correlation between Amount Purchased in Shop 1 and Distance to shop 1.

Since the scatterplot appears to be skewed, there is no strong correlation between Amount Purchased in Shop 1 and Average Price in shop 1.



There is a strong positive linear correlation between Amount Purchased and Unique products purchased in Shop1.

2. [5 points] Export the above dataset restricted to shop 1 and shop 2 shoppers to Excel. Perform a linear regression (in Excel or in other programs of your comfort. If you need to learn how to model in Excel, please feel free to leverage this video) to model amount purchased in shop 1 using the following variables
    i) Average price in shop 1
    ii) Average price in shop 2
    iii) Distance ratio (i.e, distance to shop 1 / distance to shop 2)
    iv) Unique item ratio (i.e., unique products purchased shop 1 / unique products purchased shop 2)

- Submit the model output

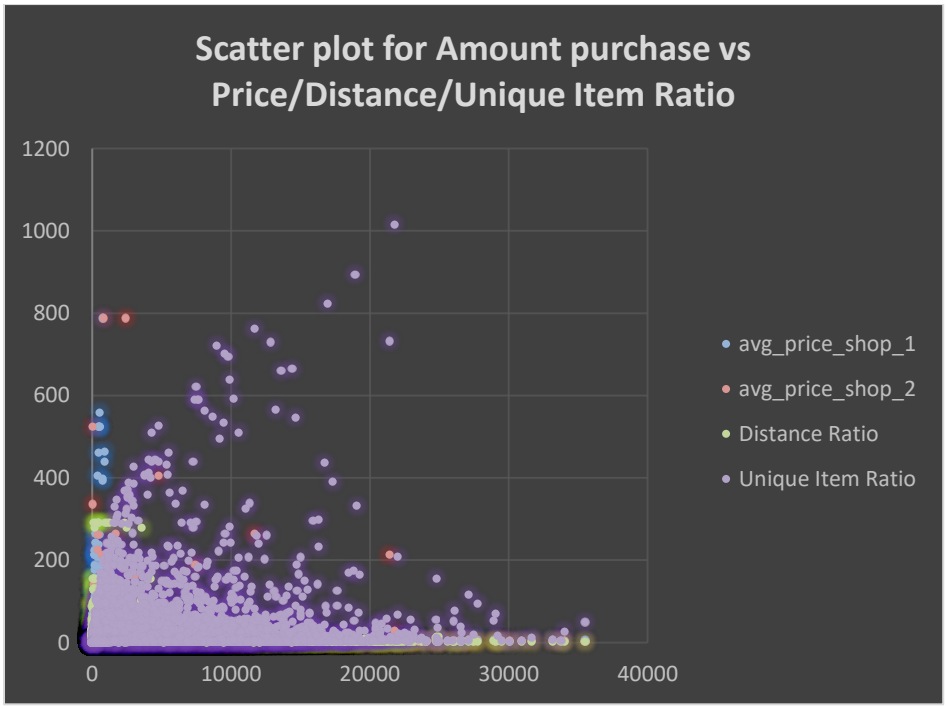
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.216139951
R Square	0.046716479
Adjusted R Square	0.046513803
Standard Error	3414.990998
Observations	18819

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	10752470984	2688117746	230.4990615	1.5349E-193
Residual	18814	2.19412E+11	11662163.52		
Total	18818	2.30164E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2275.239184	29.14966318	78.05370409	0	2218.103218	2332.37515	2218.103218	2332.37515
avg_price_shop_1	-7.109248866	1.524129754	-4.664464325	3.1155E-06	-10.09668048	-4.121817249	-10.09668048	-4.121817249
avg_price_shop_2	-11.38330338	2.274664027	-5.004388888	5.65451E-07	-15.84184978	-6.924756977	-15.84184978	-6.924756977
Distance Ratio	-13.5289964	2.366443875	-5.717015537	1.10054E-08	-18.16743957	-8.890553228	-18.16743957	-8.890553228
Unique Item Ratio	19.69055633	0.68052937	28.93417564	3.8486E-180	18.35665746	21.0244552	18.35665746	21.0244552



- What's the R-Square?

R Square 0.046716479

- What model outputs do not make sense?

  In accordance with the output model, the regression equation would be:-

  Amount Purchase in Shop1 = 2275.239-7.109*(avg_prics_shop1) -11.383*(avg_price_shop_2) -13.528(Distance Ratio) +19.69*(Unique Item Ratio)

  This implies an increase in the average price of shop 2 should decrease the amount purchased in shop 1 which seems incorrect if Shop 1 and shop 2 are competing. One would expect the Purchase amount to increase if its competitor's prices are showing a decline.

3. [5 points] We are now going to improve the model. Instead of modeling amount purchased in shop 1, we will create a new variable defined by amount purchased in shop 1 / amount purchased in shop 2. Construct a linear regression model using the following variables
   i) Price ratio (avg price shop 1 / avg price shop 2)
   ii) Distance ratio (distance to shop 1 / distance to shop 2)
   iii) Unique item ratio (unique products purchased shop 1 / unique products purchased shop 2)

- Submit the model output

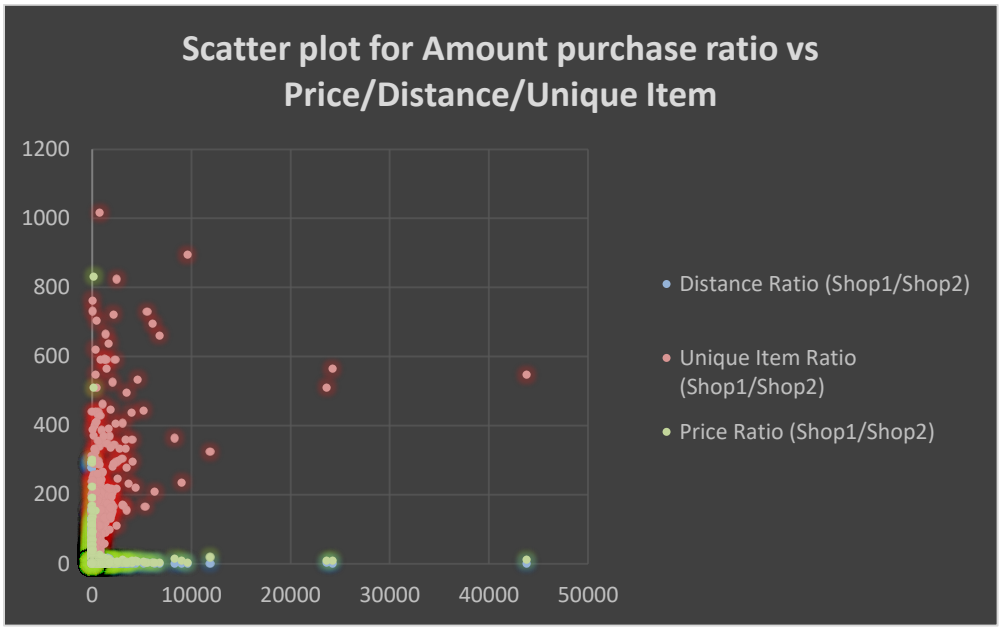
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.484934395
R Square	0.235161367
Adjusted R Square	0.235039416
Standard Error	412.8440743
Observations	18819

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	985989960.6	328663320.2	1928.320097	0
Residual	18815	3206832921	170440.2297		
Total	18818	4192822882			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-19.97126326	3.390642538	-5.890111692	3.92506E-09	-26.61722805	-13.32529846	-26.61722805	-13.32529846
Distance Ratio	0.275007167	0.286041397	0.961424359	0.336351196	-0.285659737	0.83567407	-0.285659737	0.83567407
Unique Item Ratio	6.150930682	0.081028331	75.9108645	0	5.992107853	6.30975351	5.992107853	6.30975351
Price Ratio(Shop1/Shop2)	1.167202844	0.319726817	3.650625418	0.000262311	0.540509483	1.793896205	0.540509483	1.793896205



Amount Purchase Ratio(1/2) = -19.97 + 1.167*(Price Ratio1/2) +0.275*(Distance Ratio1/2) +6.15*(Unique Item Ratio1/2)

- What's the R-Square?

  R Square 0.235161367

- Why do you think the changes we made to the model led to the R-square improvement?

  We observed that using the average price ratio instead of the independent variables significantly improved the R-square value. A good regression model can be only as good as the variables measured by the study. The results for the variables include in the analysis for Q2 can be biased by the significant variables that we don't include.

  By combining two independent predictor variables we have reduced the variance or the unnecessary noise in data that might be caused by involving too many predictors.

4. [5 points] As we are still not satisfied with the model, we will continue modifying it. Instead of modeling the purchase amount ratio (amount purchased in shop 1 / amount purchased in shop 2), we will model the log transformation of this variable. We will also apply log transformation to the set of independent variables in 3. Construct the corresponding linear regression model.
   - Submit the model output

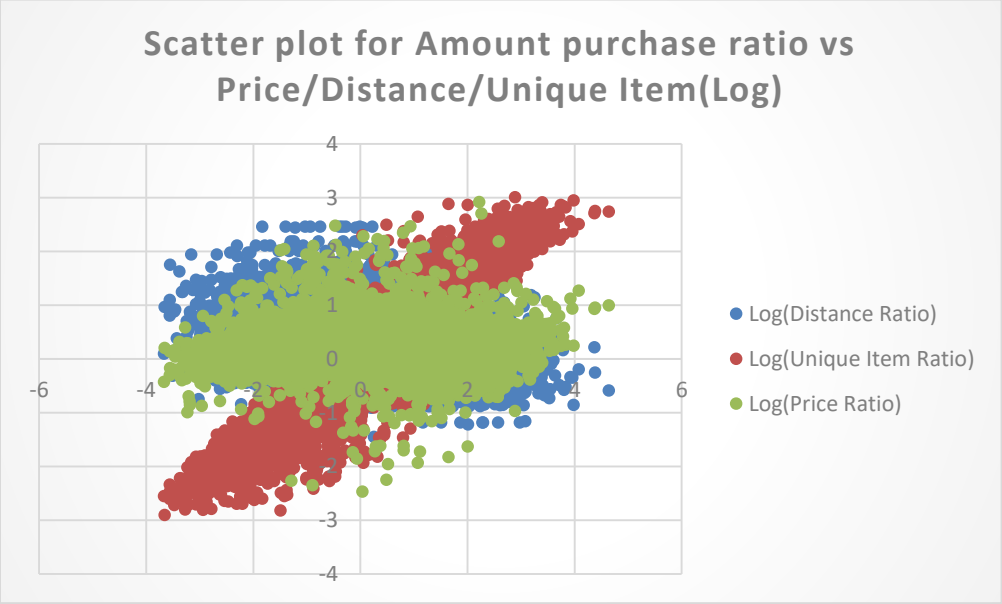
SUMMARY OUTPUT	

Regression Statistics	
Multiple R	0.98311049
R Square	0.966506236
Adjusted R Square	0.966500895
Standard Error	0.210367368
Observations	18819

ANOVA

	df	SS	MS	F	Significance F
Regression	3	24027.11743	8009.039145	180977.1184	0
Residual	18815	832.6470927	0.04425443		
Total	18818	24859.76453			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.088074324	0.00213106	-41.3288739	0	-0.092251394	-0.083897253	-0.092251394	-0.083897253
Log(Distance Ratio)	-0.07275844	0.00348465	-20.8797001	9.96313E-96	-0.079588668	-0.065928213	-0.079588668	-0.065928213
Log(Unique Item Ratio)	1.29666477	0.001992905	650.6406864	0	1.292758498	1.300571042	1.292758498	1.300571042
Log(Price Ratio)	1.015197104	0.005866577	173.047597	0	1.003698084	1.026696124	1.003698084	1.026696124



- What's the R-Square?

  R Square 0.966506236

  Log(Purchase Amount Ratio) = -0.0880 + 1.015*Log(Price Ratio) -0.07*Log(Distance Ratio) +1.296*Log(Unique Item Ratio)

- Why do you think the changes we made to the model led to the R-square improvement?

  Transformation of both the response variables and predictor x is used when the regression function is non-linear, error terms are not normal and have unequal variances.

  Hence, changing the independent and dependent variable base to natural log had a positive effect on the model and improves the R-square value as it normally distributes the data.

- Did you notice any change to the p-value associated with the distance ratio between the model in 3 and the model in 4? Why?

When running a linear regression with a log transformed response, each predicted value of ln(Y) is in natural log scale and should follow a normal distribution. Using log transformations helps in flattening the data and getting rid of unwanted variances and outliers. Hence, the p value has significantly reduced from 0.336(in Q3) to 9.96313E-96(in Q4). Such a low value of p would imply that the distance does have significant effect on the purchase amount. Increase in distance has a negative effect on purchase amount for a shop which seems plausible.

- How would you interpret the coefficients?
  Coefficient for Price ratio is positive which would mean that increase in price would also lead to increment in amount purchase which does not make sense logically.
  Coefficient for distance ratio is negative and has inverse relation with purchase amount which seems reasonable as stated above.
- Does the coefficient associated with price ratio make sense? Why do you think we end up with this type of coefficient? (Hint: think data bias)
  Coefficient of price ratio does not make sense as stated above. One would not expect an inc. in price ratio to increase the purchase amount. This suggests that there could be some form of data biasness involved. The average price ratio between shop1/shop 2 is positive which could be due to reasons such as quality offered, available brands, consumer preferences, proximity to consumers many of which are not available to us to include as part of our analysis.

5. [5 Points] Having observed the coefficient associated with the price ratio above, what can do you do to improve the model quality?
   Although the model output in Q 4 yields a higher R square value, it does not always imply that the model is good. Every time you add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases. Consequently, a model with more terms may appear to have a better fit simply because it has more terms. To improve the model quality, we can try excluding the price ratio from the output model.
   - Submit your revised model output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.955614159
R Square	0.913198421
Adjusted R Square	0.913189194
Standard Error	0.338648118
Observations	18819

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	22701.8977	11350.94885	98977.12469	0
Residual	18816	2157.866823	0.114682548		
Total	18818	24859.76453			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.083405337	0.003037122	27.46196708	8.0209E-163	0.077452305	0.089358369	0.077452305	0.089358369
Log(Distance Ratio)	-0.099503434	0.005604048	-17.75563593	5.80255E-70	-0.110487873	-0.088518995	-0.110487873	-0.088518995
Log(Unique Item Ratio)	1.256121696	0.003185919	394.2729506	0	1.249877008	1.262366384	1.249877008	1.262366384

Regression model (R- square = 0.9131)
Amount purchase Ratio = 0.083– 0.0995*(Distance Ratio) + 1.256*(Unique item Ratio)

- How does the R-square compare to your output from question 4?

R-Square value in Question 4 was 0.96 whereas for the improved model is 0.913. Although it does not improve the R value, but the dependent variables do make more sense. As stated earlier, R-squared alone cannot determine whether the coefficient estimates, and predictions are biased.

- How do you compare the applicability of this model from the previous one?
  This model removes one of the predictor variable and is a much better fit to the model. Price ratio is a parameter that was potentially biased as its values could depend on other factors that are unknown. Log Transformations for Distance Ratio and Unique Item ratio are much better predictors to study the amount purchased in shop 1/shop2.