SUMMARY
BANA 200 Foundations of Business Analytics was taught by Stephen A. Samaha, PhD for MSBA program in UC Irvine, summer 2021.

The Take Home Final includes challenges:
1. Multiple Regression in Training and Test Sample
2. Forward Variable Selection Comparison
3. Cluster Analysis and Interpretation (k-means, package: NbClust)
4. "What-If" Analysis on Increasing Variable's Value for Changing Predicted Value

---

BANA 200 Take Home Final

Dataset Description:
The cleaned text file "Starbucks.txt" contains survey data on a random sample of 6,121 Starbucks Coffee customers. The survey was done in Orange County, CA, and contains the following data:

1. **X1**: Overall, how would you rate the beverages served at Starbucks? - Taste
2. **X2**: Overall, how would you rate the beverages served at Starbucks? - Overall quality
3. **X3**: Overall, how would you rate the beverages served at Starbucks? - Temperature
4. **X4**: Overall, how would you rate the beverages served at Starbucks? - Freshness
5. **X5**: Overall, how would you rate the beverages served at Starbucks? - Presentation
6. **X6**: Overall, how would you rate the beverages served at Starbucks? - Variety
7. **X7**: Overall, how would you rate the food served at Starbucks? - Temperature
8. **X8**: Overall, how would you rate the food served at Starbucks? - Variety
9. **X9**: Overall, how would you rate the food served at Starbucks? - Taste
10. **X10**: Overall, how would you rate the food served at Starbucks? - Overall quality
11. **X11**: Overall, how would you rate the food served at Starbucks? - Presentation
12. **X12**: Overall, how would you rate the food served at Starbucks? - Freshness
13. **X13**: How do you rate the value for the money?
14. **X14**: How would you rate the Starbucks staff along the following dimensions? - Well dressed and appear neat
15. **X15**: How would you rate the Starbucks staff along the following dimensions? - Remembering your name
16. **X16**: How would you rate the Starbucks staff along the following dimensions? - Knowledgeable
17. **X17**: How would you rate the Starbucks staff along the following dimensions? - Personal treatment
18. **X18**: How would you rate the Starbucks staff along the following dimensions? - Polite
19. **X19**: How would you rate the Starbucks staff along the following dimensions? - Remembering your order correctly
20. **X20**: How would you rate the Starbucks staff along the following dimensions? - Friendly/attentive
21. **X21**: How would you rate the Starbucks staff along the following dimensions? - Have your best interest at heart
22. **X22**: How would you rate the Starbucks staff along the following dimensions? - Providing prompt service

23. **satis100**: A customer satisfaction variable that ranges from 0 to 100 points. Customers were asked the following question: “Overall, how satisfied are you with Starbucks? 0 = very dissatisfied; 100 = very satisfied.”
24. **recommend**: “How likely are you to recommend Starbucks to others? 0 = definitely WILL NOT recommend; 10 = definitely WILL recommend.” This variable ranges from 0 to 10.
25. **profits**: Average monthly profits that Starbucks earns on each customer (in US Dollars). Some profit numbers may be negative (i.e. Starbucks loses money on some customers).
26. **ZipCode**: The five digit zip code associated with the customer’s place of residence.
27. **Income**: Estimated annual income of each customer (reported in US Dollars), based on the US Census Bureau Zip Code demographics data.

Variables X1 – X22 are all measured on a 5 point scale (1 = terrible, 2 = poor, 3 = average, 4 = good, 5 = excellent).

Q1) Training and Test Samples Regression (25 Points)
Starbucks is very interested in drivers that may affect a customer’s willingness to recommend Starbucks to others. In order to help management, answer this question, do the following:

a. First divide the data into a training and test sample. Specifically, the first 5,000 observations should be the training sample, and the last 1,121 observations should be the test sample.

b. Run a multiple regression on the training sample using "recommend" as the dependent variable and X1 – X22 as the 22 independent variables. Paste the results of your regression analysis below (including all of the regression estimates and significance levels). How many of the 22 predictor variables are significant at the 5% level (have a p-value less than 0.05)? Report the R^2 value on the training sample and comment.

Result from summary(train.reg) in R

```
Call:
lm(formula = recommend ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
    X9 + X10 + X11 + X12 + X13 + X14 + X15 + X16 + X17 + X18 +
    X19 + X20 + X21 + X22, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1782 -1.3350  0.0671  1.4136  5.5228

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -3.468181   0.228723 -15.163 < 0.0000000000000002 ***
X1           0.636703   0.049315  12.911 < 0.0000000000000002 ***
X2           0.204943   0.056005   3.659             0.000255 ***
X3           0.099213   0.049019   2.024             0.043027 *
X4          -0.106475   0.048375  -2.201             0.027781 *
X5           0.230970   0.045277   5.101        0.000000349850 ***
X6          -0.047469   0.044719  -1.061             0.288520
X7           0.398114   0.047722   8.342 < 0.0000000000000002 ***
X8           0.121266   0.043533   2.786             0.005363 **
X9          -0.125841   0.050580  -2.488             0.012880 *
X10          0.350641   0.054517   6.432        0.000000000138 ***
X11         -0.080362   0.045805  -1.754             0.079418 .
X12         -0.162291   0.046437  -3.495             0.000478 ***
X13          0.381262   0.036721  10.383 < 0.0000000000000002 ***
X14         -0.158025   0.049612  -3.185             0.001455 **
X15          0.095494   0.033382   2.861             0.004246 **
X16          0.433071   0.048476   8.934 < 0.0000000000000002 ***
X17          0.065832   0.050932   1.293             0.196225
X18         -0.003917   0.051929  -0.075             0.939874
X19         -0.265705   0.047039  -5.649        0.000000017076 ***
X20          0.305544   0.054046   5.653        0.000000016610 ***
X21          0.085914   0.044156   1.946             0.051747 .
X22          0.246201   0.049649   4.959        0.000000732915 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.929 on 4977 degrees of freedom
Multiple R-squared:  0.3547,    Adjusted R-squared:  0.3519
F-statistic: 124.4 on 22 and 4977 DF,  p-value: < 0.00000000000000022
```

Table 1 Sorted P-value for Significant Variables from X1 to X22

X1	X5	X7	X10	X13	X16	X19	X20	X22	X2	X12	X14	X15
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003	0.0005	0.0015	0.0042
X8	X9	X4	X3	X21	X11	X17	X6	X18				
0.0054	0.0129	0.0278	0.043	0.0517	0.0794	0.1962	0.2885	0.9399				

Table 1 reports the sorted p-value across variables from X1 to X22, and the first 17 variables (bold) are significant with p-value less than 0.05.

The R-squared value on training sample is 0.3547, which means this model could explain 35.47% of variation depends on X1 to X22 variables.

c. Using your regression model estimated from part b) above, calculate the out-of-sample R^2 value for the 1121 observations in the test sample and report it below. Compare the R^2 value from the training sample to the R^2 value you calculated in the test sample. What can you conclude about the model's ability to predict "recommend" in the test sample? How much of a difference is there in the R^2 values between the training and test samples?

The R-squared value on test sample is 0.2944, which means this model could explain 29.44% of variation depends on X1 to X22 variables in test samples.

The difference between R-squared value on training sample and test sample is 0.3547-0.2944 = 0.0603. Our model drops 17% explanatory power on test dataset.

Q2) Variable selection (25 Points)

Using only the training sample, perform a forward variable selection procedure by using "recommend" as the dependent variable and X1 – X22 as the 22 predictor variables. Paste the results of your regression results based on the final variables selected below. Which variables were dropped? What is the R^2 of the forward selection model? When you compare the R^2 of the full model (with all 22 variables) and the R^2 of the model using forward selection, by how much did the R^2 go down by? What can you conclude about how much those dropped variables really matter?

Result from summary(forward.results) in R

```
Call:
lm(formula = recommend ~ X1 + X16 + X7 + X13 + X20 + X5 + X10 +
    X12 + X19 + X22 + X15 + X2 + X14 + X8 + X9 + X21 + X4 + X3 +
    X11, data = starbucks2)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1986 -1.3335  0.0803  1.4048  5.5491

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -3.50122    0.22527 -15.542 < 0.0000000000000002 ***
X1           0.63003    0.04901  12.854 < 0.0000000000000002 ***
X16          0.44372    0.04795   9.254 < 0.0000000000000002 ***
X7           0.39936    0.04724   8.455 < 0.0000000000000002 ***
X13          0.38447    0.03666  10.488 < 0.0000000000000002 ***
X20          0.31825    0.05078   6.267        0.000000000400 ***
X5           0.22566    0.04442   5.080        0.000000392146 ***
X10          0.34806    0.05438   6.400        0.000000000169 ***
X12         -0.16323    0.04640  -3.517              0.000440 ***
X19         -0.26808    0.04634  -5.785        0.000000007681 ***
X22          0.25119    0.04910   5.115        0.000000324831 ***
X15          0.10643    0.03240   3.285              0.001027 **
X2           0.19819    0.05585   3.548              0.000391 ***
X14         -0.15168    0.04906  -3.091              0.002003 **
X8           0.11542    0.04340   2.660              0.007847 **
X9          -0.12499    0.05055  -2.473              0.013442 *
X21          0.09116    0.04330   2.105              0.035315 *
X4          -0.10397    0.04805  -2.164              0.030514 *
X3           0.09016    0.04847   1.860              0.062952 .
X11         -0.07695    0.04571  -1.683              0.092365 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.929 on 4980 degrees of freedom
Multiple R-squared:  0.3543,    Adjusted R-squared:  0.3519
F-statistic: 143.8 on 19 and 4980 DF,  p-value: < 0.00000000000000022
```

Table 2 Comparison of Variable Used in Forward Selection Model and Full Model

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
Forward	Forward	Forward	Forward	Forward	<NA>	Forward	Forward	Forward	Forward	Forward
Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full

X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22
Forward	Forward	Forward	Forward	Forward	<NA>	<NA>	Forward	Forward	Forward	Forward
Full	Full	Full	Full	Full	Full	Full	Full	Full	Full	Full

According to table 2, X6, X17, X18 were dropped from forward selection model, using only 19 variables compared to 22 variables in full model. The R-squared value of the forward selection

model is 0.3543, which goes down by 0.0004 from 0.3547 of R-squared value from the full model.

Since there is not significant drop in R-square value between these two models, we might conclude that dropping X6, X17, X18 doesn't affect much.

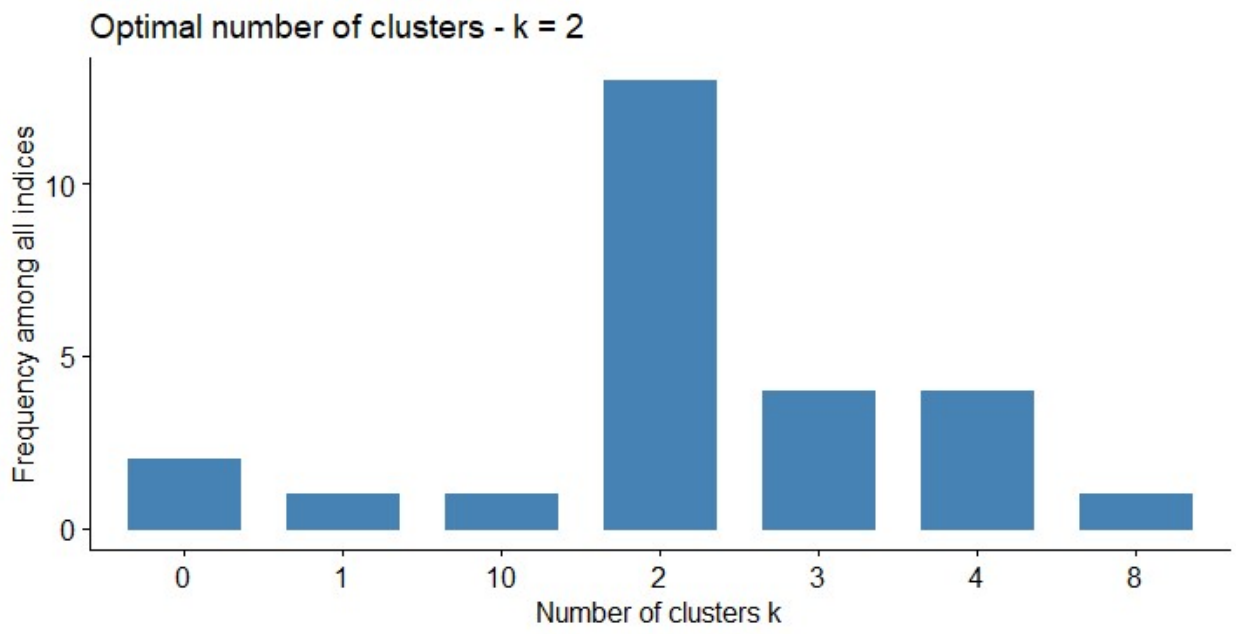
Q3) Cluster Analysis and Interpretation (25 Points)

a. Using all of the data (all 6121 observations), create a data matrix called "X" which includes the 22 predictor variables: X1, X2, …, X24, X25. Your data matrix X should have 6121 rows and 22 columns.

b. Once you have created your data matrix, use the NbClust procedure discussed in class to determine the optimal number of segments (clusters) for X. Use Euclidean distance as the distance measure, the minimum number of clusters to test = 2, and the maximum number of clusters to test = 10. Make sure to specify method = "kmeans". Based on the analysis performed, what are the optimal number of clusters for X? Use the "majority rule" to determine the optimal number of clusters. Paste the bar chart you obtained from the analysis in R below and report the optimal number of clusters to use. Note: It might take several minutes for the analysis to run, as it is computationally intensive…

According to the majority rule, the best number of clusters is 2.

Chart 1 Bar Chart of Proposed Number of Clusters



c. Using the optimal number of clusters you found in part Q3b above, run a k-means cluster analysis on the X matrix (perform a k-means cluster analysis on the X matrix using X1 – X22). Set "centers =" to the optimal number of clusters you found in step Q3b, and set the iter.max = 1000 and nstart=100. Report below how many customers are in Cluster 1 and how many customers are in Cluster 2.

Table 3 Number of customers in Cluster 1 and Cluster 2

Cluster 1	Cluster 2
3230	2891

d. Executive management has asked you to identify the "most satisfied" segment of customers. Examine the cluster centers from your k-means analysis and identify the segment of customers that seem the most satisfied. Hint: The most satisfied segment should be the one that generally has the highest average ratings (the highest cluster center values for X1 – X22). Once you have identified the most satisfied segment of customers, flag this segment and set them aside. Report below the cluster center values for X1, X2, X3, X4, and X5 (rounded to two decimal places) for this most satisfied segment of customers.

Table 4 Cluster Center Values for X1 to X5

X1	X2	X3	X4	X5
4.13	4.16	4.22	4.17	4.24

e. Executive management wants to know by how much more the "most satisfied" segment you have identified in Step 3d above is willing to recommend Starbucks as compared to all other customers. In order to answer this question, do the following:

1. Split your overall data sample into two groups: "Most Satisfied" and "All Other". The "most satisfied" group of customers should consist of the one segment that is most satisfied based on Step 3d above, and "All Other" customers will include all other customers that are not in the "most satisfied" segment.

2. Next, run two separate regression analyses for each group. Use "recommend" as the dependent variable for both regressions and X1 – X22 as the 22 predictor variables. Just to be clear: You are running two separate regressions for each one of the two groups: One regression for "Most Satisfied" and a separate regression for "All Other".

Result from summary(most_satisfied.reg)in R

```
Call:
lm(formula = recommend ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
    X9 + X10 + X11 + X12 + X13 + X14 + X15 + X16 + X17 + X18 +
    X19 + X20 + X21 + X22, data = most_satisfied)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0736 -1.2631  0.1512  1.4830  4.5561

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) -1.92561    0.50841  -3.787             0.000155 ***
X1           0.59117    0.06287   9.403 < 0.0000000000000002 ***
X2           0.26784    0.06958   3.849             0.000121 ***
X3           0.10633    0.06090   1.746             0.080894 .
X4          -0.21958    0.06151  -3.570             0.000362 ***
X5           0.14806    0.05776   2.563             0.010414 *
X6          -0.08490    0.05580  -1.521             0.128239
X7           0.34366    0.05928   5.797     0.00000000739016 ***
X8           0.10838    0.05368   2.019             0.043578 *
X9          -0.11867    0.06156  -1.928             0.053992 .
X10          0.31133    0.06597   4.719     0.00000246963561 ***
X11         -0.12390    0.05710  -2.170             0.030092 *
X12         -0.07768    0.05740  -1.353             0.176055
X13          0.29807    0.04558   6.539     0.00000000007183 ***
X14         -0.14176    0.06327  -2.241             0.025114 *
X15          0.08476    0.04298   1.972             0.048687 *
X16          0.42770    0.06060   7.058     0.00000000000206 ***
X17          0.06970    0.06517   1.070             0.284906
X18         -0.01037    0.06757  -0.153             0.878073
X19         -0.19696    0.06033  -3.265             0.001108 **
X20          0.24661    0.06869   3.590             0.000336 ***
X21          0.07159    0.05440   1.316             0.188232
X22          0.21803    0.06256   3.485             0.000499 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.892 on 3207 degrees of freedom
Multiple R-squared:  0.1622,    Adjusted R-squared:  0.1564
F-statistic: 28.22 on 22 and 3207 DF,  p-value: < 0.00000000000000022
```

Result from summary(all_other.reg)in R

```
Call:
lm(formula = recommend ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
    X9 + X10 + X11 + X12 + X13 + X14 + X15 + X16 + X17 + X18 +
    X19 + X20 + X21 + X22, data = all_other)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9145 -1.4339  0.0217  1.3916  5.7327

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -3.34446    0.43682  -7.656   0.0000000000000260 ***
X1           0.65182    0.06555   9.944 < 0.0000000000000002 ***
X2           0.28766    0.07512   3.829              0.000131 ***
X3           0.03067    0.06655   0.461              0.644903
X4          -0.12505    0.06424  -1.947              0.051686 .
X5           0.25323    0.05914   4.282   0.0000191348748606 ***
X6          -0.02377    0.05926  -0.401              0.688334
X7           0.49032    0.06405   7.656   0.0000000000000261 ***
X8           0.09597    0.06017   1.595              0.110794
X9          -0.19581    0.07069  -2.770              0.005644 **
X10          0.37227    0.07542   4.936   0.0000008418275368 ***
X11          0.02180    0.06219   0.351              0.725975
X12         -0.25048    0.06405  -3.911   0.0000941572470839 ***
X13          0.46570    0.05002   9.310 < 0.0000000000000002 ***
X14         -0.14464    0.06547  -2.209              0.027233 *
X15          0.08864    0.04448   1.993              0.046389 *
X16          0.40606    0.06454   6.292   0.0000000003618032 ***
X17          0.09366    0.06668   1.404              0.160279
X18         -0.05609    0.06851  -0.819              0.412978
X19         -0.24164    0.06068  -3.982   0.0000699369162700 ***
X20          0.27859    0.07071   3.940   0.0000835374046899 ***
X21          0.07423    0.06006   1.236              0.216542
X22          0.11475    0.06622   1.733              0.083231 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.006 on 2868 degrees of freedom
Multiple R-squared:  0.2122,    Adjusted R-squared:  0.2062
F-statistic: 35.12 on 22 and 2868 DF,  p-value: < 0.00000000000000022
```

3. For each one of the regressions, report the average predicted values. That is, extract the two sets of predicted values from the two lm objects by using the “fitted.values” function, and for each regression, take the average of these fitted values and report these two averages below. By how much more (in terms of average predicted “recommend”) is the “Most Satisfied” segment likely to recommend Starbucks? Round all answers to two decimal places.

Table 4 Average Predicted Recommend Values from Segment “Most Satisfied” and “All Other”

Most Satisfied	All Other
7.33	5.19

The average predicted recommend values from the “Most Satisfied” segment is 2.14 more than the values from the “All Other” segment.

Q4 "What-If" Analysis (25 Points)

Management wants to figure out by how much more it can increase the "All Other" segment's willingness to recommend Starbucks to others. It has conducted some market research and plans to invest in a series of advertisements. Based on the preliminary market research, management believes that it can increase each customer's ratings in the "All Other" segment by one point for X1, X2, X7, X8, and X10. Starbucks has asked you to recalculate the average willingness to recommend for the "All Other" segment if each customer's survey ratings increases by 1 points for X1, X2, X7, X8, and X10 in that segment.

This question is asking you to do the following:

a. For the "All Other" segment only, increase X1, X2, X7, X8, and X10 by one point. For example, if Customer 1 has X1 = 3, you should set X1 = 4 for this customer. However if Customer 1 has X1 = 5, you should NOT change his or her rating. Remember: the surveys are on a 5 point scale so you should not have any ratings above a "5".

b. Once you have changed the ratings for X1, X2, X7, X8, and X10 for the "All Other" segment, only use your existing regression model results from Q3 to recalculate the predicted "recommend" for the "All Other" segment. So, recalculate the predicted "recommend" for the "All Other" segment but make sure to use the new values for X1, X2, X7, X8, and X10 as the basis for these predictions. Do NOT rerun your regression analysis: Use the existing regression results to recalculate "recommend".

c. Once you have recalculated the predicted values for all customers in the "All Other" segment (based on the updated values for X1, X2, X7, X8, and X10), take the average of these new predicted values and report this average below, rounded to two decimal places. By how much is the willingness to recommend expected to increase by if Starbucks can get the "All Other" customer segment to be one point more satisfied for X1, X2, X7, X8, and X10? Does this seem like a worthwhile thing to do? Comment on whether the change seems significant or not.

Table 5 Average Predicted Recommend Values from "All Other" and "All Other_adjusted"

All Other	All Other_adjusted
5.19	7.05

The average predicted recommend values for the "All Other_adjusted" increased from 5.19 to 7.05, almost 36% growth, which is considered significant. In other word, except those most satisfied customers, it is worthwhile for Starbucks to increase one point in X1, X2, X7, X8, and X10 to earn estimated more 36% word-by-mouth marketing from customers.