CS 59970 - Intro to Data Science
Midterm Exam
November 11, 2019

Answer Key

This is a 90 minute exam. Please use your time wisely.

Multiple Choice:	___ / 70
Short Answer:	___ / 66
Total:	___ / 136

Academic dishonesty is prohibited in The City University of New York. Penalties for academic dishonesty include academic sanctions, such as failing or otherwise reduced grades, and/or disciplinary sanctions, including suspension or expulsion.

Part I: Multiple Choice. (2 points per question)

Circle the best answer(s). Questions that indicate "select all that apply" may have more than one correct answer. For all problems, one point will be awarded for correctly identifying one correct answer, and a second for correctly identifying all correct answers.

1. Data science modeling problems are best summed up by which of the following equations:
   a) slides w/ gifs > slides w/o gifs
   b) stats + programming + magic = data science
   c) argmax(cuny) = ccny
   d) *f* (features) = target

2. In the first lecture, we discussed a famous case in which Target [*answer below*], an example of a ______ problem.
   a) suggested complimentary groceries to customers browsing for household items
   b) forecast the total spend for each visitor to their store
   c) correctly predicted a subset of their customers who were pregnant
   d) correctly identified which online shoppers would later buy items in-store

3. In the first lecture, we discussed a famous case in which Target _______, an example of a [*answer below*] problem.
   a) regression
   b) unsupervised learning
   c) recommendation system
   d) classification

4. As discussed in class, ETL stands for:
   a) Exchange, Transact, Liquidate
   b) Extract, Transform, Load
   c) Evaluate, Teach, Learn
   d) Engineer, Translate, Learn

5. Which of the following best describe *structured* data? *Select all that apply.*
   a) Data that has a well-defined model or schema.
   b) Data that has been subject to very little preprocessing.
   c) Data with more columns than rows.
   d) Data that is clearly organized and well documented.

6. Which of the following are ***non linear*** models? *Select all that apply.*
    a) Linear regression
    b) Random forest regressor
    c) Logistic regression
    d) Decision tree classifier

7. Which of the following best summarizes the functionality of Airflow?
    a) Airflow is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models.
    b) Airflow a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.
    c) Airflow is an open-source, distributed, versioned, non-relational database modeled after Google's Bigtable.
    d) Airflow is a platform to programmatically author, schedule and monitor workflows.

8. Which of the following is a method of *handling outliers* that we discussed in class. *Select all that apply.*
    a) Trim
    b) Drop
    c) Be Mindful
    d) Impute

9. Which of the following are examples of normalizing data? *Select all that apply.*
    a) Scaling data observations so that each column has the same standard deviation.
    b) Subtracting the mean from each column of a sample of data such that each feature has mean zero and is more comparable to each other.
    c) Dropping null values so that each row in a data frame has valid values.
    d) Splitting the data into test and train sets.

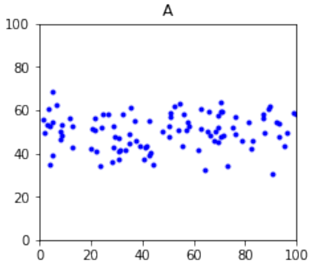
10. Which of the following are things we can learn from NYC restaurant inspection data? *Select all that apply.*
    a) The number of restaurants cited for filth flies.
    b) The number of Dunkin Donuts locations in New York City.
    c) The mean price of items at NYC restaurants.
    d) The mean health department rating by cuisine type.

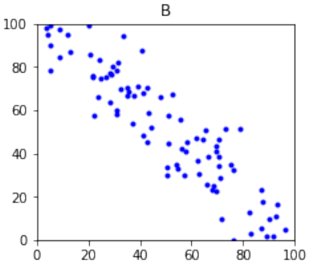
11. Which definition below best represents the textbook definition of statistics?
    a) The semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining).
    b) The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.
    c) The scientific study of algorithms and models to perform a specific task without using explicit instructions, relying on patterns and inference instead.
    d) The quantitative analysis of actual phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference.

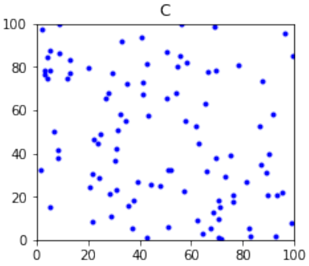
12. Which of the following are - by themselves, without additional data points - measures of the central tendency of a statistical distribution? *Select all that apply.*
    a) Median
    b) Mean
    c) Standard Deviation
    d) Range

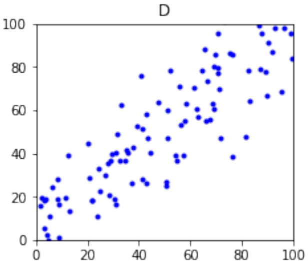
13. Which of the following statements are true of linear regression ***but not*** logistic regression.
    a) The target is a continuous variable.
    b) The target is a binary variable.
    c) The features may contain continuous variables.
    d) The model is linear in form.

14. Which of the following images below appear to show a correlation between the data plotted on the X and Y axes? *Select all that apply.*









a) Figure A
b) Figure B
c) Figure C
d) Figure D

15. Which of the following are likely to be normally distributed? *Select all that apply.*

a) The residual error terms from a well-designed linear regression.
b) The means of repeated samples of 30 individual observations drawn from an exponentially distributed population.
c) The results of individual rolls of a six-sided fair die.
d) The outcomes of individual flips of a fair coin.

16. Which option best describes the formula for calculating the probability of successes for the binomial distribution below?

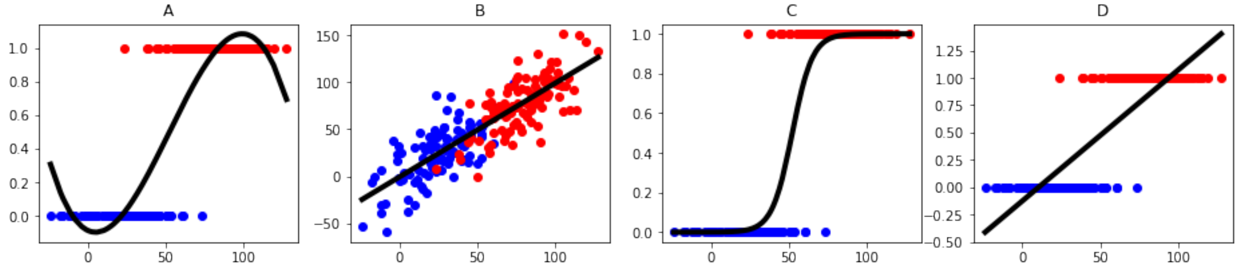
$$P(D = b) = \binom{a}{b} c^b (1 - c)^{(a-b)}$$

a) The likelihood for ***a*** successes over ***b*** trials with ***c*** probability of success.
b) The likelihood for *b* successes over *a* trials with *c* probability of success.
c) The likelihood for ***c*** successes over ***a*** trials with ***b*** probability of success.
d) The likelihood for ***c*** successes over ***b*** trials with ***a*** probability of success.

17. Which of the following are assumptions of linear regression? *Select all that apply.*

a) The errors are normally distributed.
b) The errors are uniformly distributed.
c) The error terms have a mean of zero.
d) The error terms have a standard deviation of one.

18. Which of the follow figures below depicts the output of logistic regression?



a) A
b) B
c) C
d) D

19. Which of the following are assumptions of linear regression? *Select all that apply.*

a) The independent variables (features) are independent of each other.
b) The independent variables (features) are continuous.
c) The independent variables (features) are normally distributed.
d) A linear relationship exists between the independent variables (features) and the dependent variable (target).

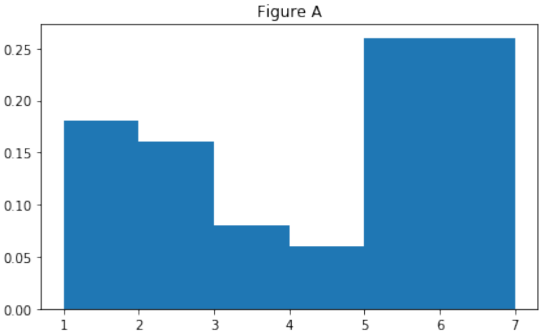
20. Which of the following are examples of continuous data? *Select all that apply.*

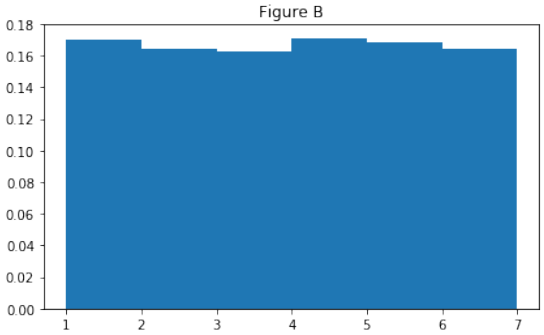
a) The age of current Netflix subscribers.
b) The level of education (*ie: high school, college, advanced degree*) of current Netflix subscribers.
c) The occupational classifications of current Netflix subscribers.
d) The incomes of current Netflix subscribers.

21. The median age of commuters traveling through Grand Central is 41 years old. A group of students randomly survey the ages of travelers passing through Grand Central. When they meet at the end of the day, they calculate the mean of each of their samples and find that the means appear to be normally distributed around 41. Which of the following best predicts this outcome?

a) The law of large numbers.
b) The law of total variance.
c) The central limit theorem.
d) Slutsky’s theorem.

22. Anjali is a newly hired data scientist at a major casino, and wants to make sure each die the casino uses is fair (equally weighted). To do so, she rolls a randomly selected die 50 times and plots the results in the histogram in Figure A. She's not sure if this really shows that the die is fair, so she rolls it another 5,000 times and plots the results in the histogram in Figure B, essentially confirming that each of the outcomes on the die is equally likely. Which of the following best predicts this outcome?





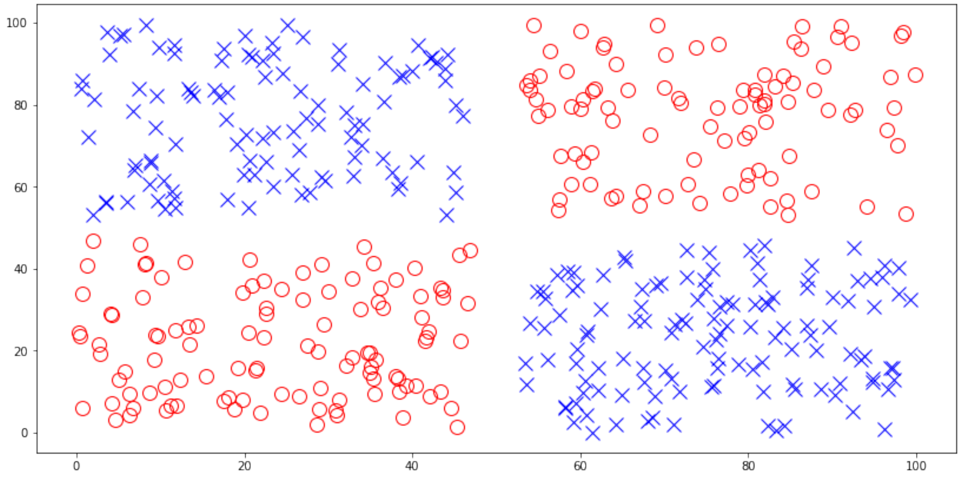
a) The law of large numbers.

b) The law of total variance.

c) The central limit theorem.

d) Slutsky's theorem.

23. Which of the following models is capable of perfectly modeling the Xs and Os in the plot shown below? *Select all that apply.*



a) Logistic regression

b) Linear regression

c) Gradient boosting machine

d) Decision tree

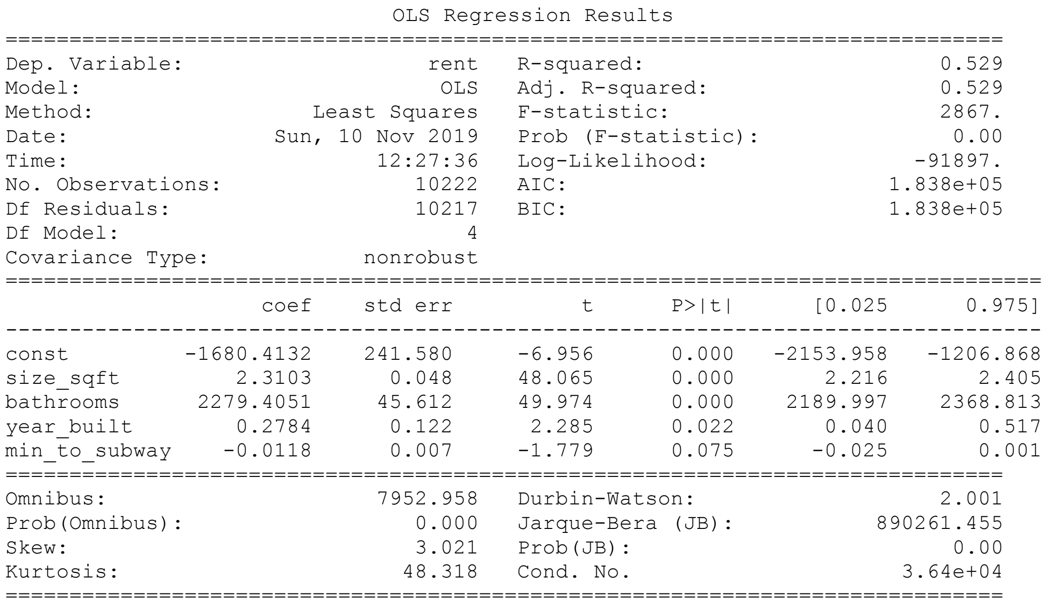
24. Which of the following are valid performance measures for a *classification* model? *Select all that apply.*
    a) R-Squared
    b) Area Under the Curve (AUC)
    c) Precision
    d) Mean Squared Error (MSE)

25. In statistics, a sequence or a vector of random variables (features) exhibits multicollinearity if:
    a) the error terms from the random variables (features) exhibit autocorrelation.
    b) one predictor variable (feature) can be linearly predicted from the others with a substantial degree of accuracy.
    c) the dependent variable (target) can be linearly predicted from the independent variables (features) with a substantial degree of accuracy.
    d) all variables (features) in the sequence or vector are drawn randomly.

26. Which of the following is an example of a nonlinear relationship between two features:
    a) When predicting rent, number of bedrooms is highly correlated with square footage of a home.
    b) When predicting survival rates for the Titanic, older women were more likely to survive than younger women, and older men were more likely to die than younger men.
    c) When predicting flight delays, each flight can have at most one origin and one destination.
    d) When predicting which Kiva loans were likely to be funded, including the *funded_amount* as a feature would result in a perfect prediction.

27. Which of the following are hyperparameters used to limit overfitting of a decision tree? *Select all that apply.*
    a) Maximum tree depth.
    b) Minimum standard deviation per leaf.
    c) Maximum entropy per split.
    d) Minimum number of observations per leaf.

Figure



```
                              OLS Regression Results
==============================================================================
Dep. Variable:                   rent   R-squared:                       0.529
Model:                            OLS   Adj. R-squared:                  0.529
Method:                 Least Squares   F-statistic:                     2867.
Date:                Sun, 10 Nov 2019   Prob (F-statistic):               0.00
Time:                        12:27:36   Log-Likelihood:                -91897.
No. Observations:               10222   AIC:                         1.838e+05
Df Residuals:                   10217   BIC:                         1.838e+05
Df Model:                           4
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const         -1680.4132    241.580     -6.956      0.000   -2153.958   -1206.868
size_sqft         2.3103      0.048     48.065      0.000       2.216       2.405
bathrooms      2279.4051     45.612     49.974      0.000    2189.997    2368.813
year_built        0.2784      0.122      2.285      0.022       0.040       0.517
min_to_subway    -0.0118      0.007     -1.779      0.075      -0.025       0.001
==============================================================================
Omnibus:                     7952.958   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           890261.455
Skew:                           3.021   Prob(JB):                         0.00
Kurtosis:                      48.318   Cond. No.                     3.64e+04
==============================================================================
```

28. Based on the figure above, which of the following numbers below best reflects the overall goodness of fit of the model?
    a) -1680.4132
    b) 0.529
    c) 241.580
    d) -6.956

29. Based on the figure above, what is the percentage likelihood that the coefficient for **`year_built`** is equal to zero?
    a) 0.000
    b) 0.122
    c) 0.2784
    d) 0.022

30. Based on the figure above, how much would monthly rent increase with a ***10 square foot increase*** in total square footage of an apartment with one bathroom that is seven minutes away from the subway?
    a) $0.48
    b) $2,279.41
    c) $-0.08
    d) $23.10

31. Which of the following is best described as a plot of the true positive rate against the false positive rate at every possible threshold from highest to lowest?
    a) Lift curve
    b) Indifference curve
    c) Receiver operating curve
    d) Efficient frontier

32. In statistics, a sequence or a vector of random variables is homoskedastic if:
    a) all variables in the sequence or vector have increasing variance.
    b) Each observation in the sequence or vector has variance proportional to the corresponding error term.
    c) all variables in the sequence or vector are normally distributed.
    d) all variables in the sequence or vector have the same finite variance.

33. Which of the following are valid performance measures for a *regression* model? *Select all that apply.*
    a) R-Squared
    b) Area Under the Curve (AUC)
    c) Precision
    d) Mean Squared Error (MSE)

34. Which of the following are examples of ensemble models? *Select all that apply.*
    a) Random forest
    b) Decision tree
    c) Gradient boosting machine
    d) Logistic regression

35. Which of the following are examples of data scientists applying the principle of Occam's Razor? *Select all that apply.*
    a) Akshay forces his model to use just the 15 most predictive features out a possible 110 input to the model training process.
    b) Brandi decides to deploy a simple model that performs well over a more complex model the performs marginally better on her training data.
    c) Christian uses one-hot-encoding for a single categorical feature, resulting in 60 different boolean features for his linear regression.
    d) Donya uses a convolutional neural net to forecast next year's economic growth.

Part II: Short Answer.

Answer each of the following in a few short sentences.

1. **14 points.** Wei is a data scientist at Hinge, a dating app, and has built a model to predict whether couples that went on a first date in the first week of September will go on a second date or break up. After training his model, he applies it to a holdout sample of 100 couples and predicts that 30 couples will go on second dates. However, he finds that of the 30 couples he predicted would go on second dates, only 20 did and 10 broke up. Moreover, he realizes that in total, 40 of the couples in the sample of 100 actually did go on second dates, and the other 60 broke up.

   a) Complete the confusion matrix for Wei's model. (6 pts)

		Predicted		
		Go on 2nd Date ❤️	**Break Up** 😢	**Total**
Actual	**Go on 2nd Date** ❤️	**20**	**20**	40
	Break Up 😢	**10**	**50**	**60**
	Total	30	**70**	100

Continued on next page.

b) What is the accuracy of Wei's model? (2 pts)

(20 + 50) / 100 = 70%

c) What is the recall of Wei's model? (2 pts)

20 / (20 + 20) = 50%

d) Wei predicted 30 couples would go on second dates, but 40 actually went on second dates. Without retraining his model with new or better data, what could Wei do to predict that more couples would go on second dates? Do you think this would cause recall to go up, down, or stay the same? Why? (4 pts)

Wei should lower threshold of predictive model so that more couples are predicted to go on a second date. Predicting more couples will go on a second date should raise recall, or at least keep recall the same in the case that all of the newly predicted couples break up.

2. **12 points.** Alexei is an applied scientist at Amazon, and is building a forecasting model to predict **Y** based on **X1**, **X2**, and **X3**. His data is shown in the table below.

Obs.	X1	X2	X3	Y
1	A	0.73	1	26
2	B	0.79		54
3	A	0.68	2	25
4	B	0.97		70
5	A	0.67	2	28
6	A	0.60		10
7	B	0.45	4	52
8	B	0.95	4	18

a) Alexei has observations with missing data for **X3**. Give Alexei **three** strategies for handling the missing values, a brief advantage and disadvantage for pursuing this strategy, along with the resulting values for observations 2, 4, and 6 for **X3** (if appropriate). (9pts)

Alexei could drop the affected rows, leaving fewer observations, or drop the affected column, leaving less predictive info for the model. He could also impute the values with the mean (2.6) or median (3), or modeled values (noticing that X3 is 1 or 2 when X1=A and 4 when X2=B, he could set #2 and #4 equal to 4 and #6 equal to 1). This would leave all columns and rows intact, but potentially introduce new errors.

b) Alexei's colleague Rob suggests using a logistic regression to predict **Y**. Is Rob's suggestion a good one? Why or why not? (3 pts).

Rob's suggestion is bad; this is a regression problem and logistic regression is a classification algorithm.

3. **8 points.** Lauren is an applied scientist at Lyft, a ride-sharing platform, studying the New York market, and wants to know how many drivers Uber, a rival ride-sharing platform, has signed up. Fortunately for Lauren, Uber has introduced a new feature that displays the driver number each time you use Uber, and this driver number has been sequentially assigned, meaning that the first driver who signed up was #1, the second #2, the hundredth #100. Lauren assumes the likelihood of seeing any particular driver is uniformly distributed, and has 50 interns call Ubers for a day and record the driver numbers, and has a data set of 250 driver numbers.

   a) Based on the all of the driver numbers she has seen, how should Lauren calculate her best guess for the total number of drivers on the Uber platform? (4 pts)

   Good guesses include mean x 2, the max, or the max + max / no. of obs.

   b) What is the advantage of this strategy? What is the potential downfall of this strategy? (4 pts)

   Each strategy balances some elements of bias (the max will always be below the actual value) and variance (when n is small, the mean x 2 can vary widely and may not converge smoothly). The max + max / n strategy is optimal, but also may be subject to error given the low sample size.

4. **6 points.** Jorge is doing policy research for the Mayor's office and has two data sets: the incomes of all New Yorkers, and the heights of all New Yorkers, and he wants to know if he can assume that either of these data sets is normally distributed.

   a) Which of these two data sets is most likely to be normally distributed? For each data set, give a reason why it may or may not be normally distributed. (4 pts)

      Generally height is more likely to be normally distributed: there are a lot of people of average height, some short people and some tall people, but the number of each other those should be roughly equal.

      Income is very likely to be skewed upward, with many people making below the mean, and a long tail of rich individuals making a very high amount.

      ***(A variety of answers were ultimately accepted)*.**

   b) What might Jorge do to determine whether these data sets are normally distributed? (2 pts)

      Jorge could plot a histogram of his data, or take other steps to see if it is skewed (ie. calculate the mean and median).

5. **12 points.** Dante is a data science intern at Google, and has received a call from Fang, a product manager at Youtube. Fang wants Dante to create a model to predict the likelihood that a user who has watched a given publisher video will watch another video by the same publisher immediately after.

   a) Name a suitable model for solving this problem. Why would you choose this model? (4 pts)

      This is a classification model, so logistic regression, decision tree, random forests, or gradient boosting machines would work.

   b) Give an example of **two** features that might be helpful in predicting the target along with a short reason why you think this might help Dante solve the problem. (4 pts)

      A number of things could work here, but subscription status, video popularity, video content, and user demographics are obvious choices.

   c) Dante has 30,000 observations to train his model but is concerned he might be overfitting. What might Dante do to ensure he does not overfit? How can he obtain a reliable measurement of his performance on unseen data? (4 pts)

      Dante should select hyperparameters to limit his model's ability to overfit, and use cross validation, or at least a hold out set to estimate performance on unseen data.

6. **14 points.** Natalia is an quantitative analyst at Starbucks and has been tasked with estimating the total amount each customer that visits a Starbucks location during the last week in November will spend during the month of December.

   a) Is this a classification or regression problem? Why? (2 pts)

      Total spend is a continuous target, so this is a regression problem.

   b) Name a suitable model for solving this problem. Why would you choose this model? (4 pts)

      Linear regression is an obvious choice, though a decision tree regressor (or random forest or gradient boosting regressor) could also work.

   c) Give an example of **two** features that might be helpful in predicting the target along with a short reason why you think this might help Natalia solve the problem. (4 pts)

      The number of visits in the period by the customer, the items they order, prices at a given location, how far customers travel, weather patterns, and holidays could all be helpful predictors.

   d) How might Starbucks want to use this model, and in what way might this model be wrong, but useful? (4 pts)

      Estimates of total spend could help Starbucks in planning, staffing, targeting stores for additional ad spend, and managing future cash flows and earnings expectations. It's highly unlikely any model would predict this perfectly, but a model with decent predictive power could be very helpful in the aforementioned tasks.