Problem 1: A large auto dealership is interested in determining the number of cars that will be sold in a given week. The management of the dealership believes that a relationship can be found between the number of cars sold (Y), the advertised price (X1) and the current interest rates (X2). Their past experience shows that they tend to have better luck using a non-linear relationship. **The below model shows the result when using Y (i.e. the number of cars sold) as the response variable. Note that both price and interest are log-transformed; the number of cars sold is not transformed.**

Summary measures

R-Square	0.8698
Adj R-Square	0.8498

Regression coefficients

	Coefficient	*Std Err*	*t-value*	*p-value*
Constant	4.3965	0.7549	5.8239	0.0001
Log Price	-0.0825	0.2467	3.3456	0.0053
Log Interest	-0.1225	0.1880	-0.6512	0.5262

a. Consider the number -0.0825 in the column labeled "Coefficient." Provide a precise economic interpretation (i.e. one that your boss or client would understand) for this number.

   i. If price increases by $1 (and interest rates remain constant), the number of cars sold decreases by 0.0825 units.
   ii. If price increases by $1 (and interest rates remain constant), the number of cars sold increases by 0.0825 units.
   iii. If price increases by 1% (and interest rates remain constant), the number of cars sold decreases by 8.25%
   iv. If price increases by $1 (and interest rates remain constant), the number of cars sold decreases by 0.0825%
   v. If price increases by 10% (and interest rates remain constant), the number of cars sold decreases by 0.00825 units.
   vi. None of the above.

Note that the response is original scale but the predictor is logged; hence the decrease in cars is (0.1)(0.0825) = 0.00825 units

b. Based on the above model, what is the predicted demand (i.e. number of units sold in a week) for an advertised price of $15,000 and an interest rates of 5%? (For the interest rate, use 0.05 in all your calculations!)

   i. The predicted demand is approx. 0 cars.
   ii. The predicted demand is approx. 1 car.
   iii. The predicted demand is approx. 3 cars
   iv. The predicted demand is approx. 4 cars
   v. The predicted demand is approx. 5 cars
   vi. The predicted demand is more than 10 cars

The answer is 3.97. Hence the demand is closest to 4.

c. Based on the above regression model, the relationship between predictors and response can best be characterized as

   i. Weak
   ii. Linear
   iii. Quadratic
   iv. Polynomial
   v. None of the above

The relationship is logarithmic.

Problem 2: Consider the data “Spam.csv” posted on Canvas. For these data, estimate a logistic regression model with “Spam” as the response variable and the following 4 predictor variables: “lab”, “conference”, “credit” and “money”. Based on that model, answer the following questions below:

a. What is the AIC value for this model?
   i. 1932
   ii. 1455
   iii. 1230
   iv. 1021 – The AIC is 1021.6
   v. None of the above.
b. Which variable is the least significant predictor?
   i. Lab
   ii. Conference – has a p-value of only o.o5342
   iii. Credit
   iv. Money
c. Which of the following statements is true?
   i. All else equal, “Credit” increases the odds of Spam by a factor of less than 1,000. – No, by a factor of more than 1,000 – see below.
   ii. There is a positive relationship between “Lab” and Spam. – Negative!!
   iii. All else equal, for each additional occurrence of the word “Conference”, the probability of Spam decreases by 5.3392. – No, odds decrease by exp(5.3392)
   iv. All else equal, “Money” increase the probability of Spam by a factor or more than 1,700. – No, it increases the odds, not probability.
   v. None of the above. – Coef of credit is 7.4738; hence odds increase by exp(7.4738) = 1761.287, a factor more than 1,000.

Problem 3: Consider the data “AutomobileMarket.csv” posted on Canvas. Cluster these data, using the k-means algorithm with k=2 clusters. (Use all the defaults.) Based on that clustering, answer the following questions.

a. For the first cluster, compute the mean score for all 10 cars in the data (i.e. the mean score for BMW328i, the mean score for Ford Explorer, and so on). Repeat the same for cluster two. Comparing the mean sores in cluster 1 with those in cluster 2, which car has the largest change (i.e. absolute difference) in mean sore?
   i. BMW328i
   ii. Infinity J30
   iii. Mercedes C280
   iv. Porsche Boxster – Mean in C1 = 4.426; mean in C2 = 6.439
   v. Volvo V90
b. Next, compute the correlation between the scores for BMW328i and the scores for Ford Explorer. Compute the correlation both for cluster 1 and cluster 2. Based on these two correlation values, we can conclude that…
   i. …BMW328i and Ford Explorer have a positive relationship both in cluster 1 and in cluster 2.
   ii. …BMW328i and Ford Explorer have a negative relationship both in cluster 1 and in cluster 2.
   iii. …BMW328i and Ford Explorer have a negative relationship in cluster 1 but a positive relationship in cluster 2.
   iv. None of the above.

Problem 4: Consider the data “Enrollments.csv” posted on Canvas. Run a simple regression model with enrollment (“Roll”) as the response variable and lag-1 of enrollment as the (only) predictor. To that end, you first have to create a lag variable of order 1 for enrollment. Note that after creating the lag variable, the size of the resulting data set will reduce by one row. Based on that model, answer the following questions below:

a. What is the value of R-squared for that model?
   i. 28%
   ii. 48%
   iii. 58%
   iv. 78%
   v. 98% - R-sq equals 0.9842

b. Consider the coefficient for the lag-1 of enrollment. Based in that coefficient, we can conclude that…
   i. The growth-rate of enrollment decreases by a factor of 0.92 in every year.
   ii. All else equal, enrollment increases by a factor of 0.92 in every year.
   iii. The rate at which enrollment increases grows by a factor of 0.92 in every year.
   iv. Lag-1 of Enrollment is not significant.
   v. None of the above.

Note that the coef of lag-1 enrollment is 0.916; hence enrollment increases by a factor of 0.916 of last years enrollment; that is, although enrollment continues to increase, the rate slowly goes down as year over year we only see 92% of last year’s enrolment increase. For instance, in year 2, enrollment is intercept + 0.92 * year1_enrollment; in year 3, enrollment is intercept + 0.92 *b year2_enrollment, or intercept + 0.92 * (intercept + 0.92 * year1_enrollment) = 1.92 * intercept + 0.92^2 year1_enrollment, and so on.