Question 1.

1A.

In this case, the price of a pre-owned Prius would be the response variable (*y1i*) and the explanatory variable would be mileage (denoted *x2i* on the exam). Age (*x1i*) is not considered here.

1B.

Using the given R output, we can see that the least squares estimate for mileage (*x2i*) is -7.109e-02. This represents the slope of the linear regression model, which predicts that the mean price of a Prius decreases by $0.07109 for every mile. The least square estimate for the y-intercept is 1.952e+04, which represents the mean price of a new Prius with zero miles ($19,520).

1C.

The four assumptions for a least squares regression analysis are (1) mean zero linearity, where the average value of errors is 0 regardless of values of any predictors of the response, (2) homoscedasticity, where variance is equally distributed around 0 on residual plot, (3) error normality, where errors are normally distributed, and (4) independence, where errors are uncorrelated. The mean zero linearity (assumption 1) is approximately zero. However, examination of the residuals median of -211.1, Q1 -2380.9, Q3 of 2002.7, and plot of “res vs miles” does show that most data points deviate slightly negatively from zero. Assumption 2 of homoscedasticity is close to being met. The “res vs miles” plot above, shows evenly distributed residuals less than ~12,500 miles, but mostly negative residuals above ~12,500 miles. Assumption 3 of error normality is close to being satisfied. The residuals median of -211.1, Q1 -2380.9, Q3 of 2002.7 all indicate an approximately normal distribution. The “res vs miles” plot shows that residuals are approximately evenly distributed between -5,000 to 5,000. The “Normal Q-Q” plot shows an approximately straight line, although there is noticeable sigmoidal deviation within the data points. Assumption 4 of independence, is met, and can be seen on plots “price vs miles” and “res vs miles” as neither plot shows any correlation between errors. Overall, the assumptions are met, although we do see at least two data points that have unusual values that should be examined further.

Question 2

2A.

Point #1 has an unusual x (miles) in that it lies well beyond the rest of the dataset. The residual of #1 is not unusual in comparison to the rest of the dataset. In this sense, Point #1 is considered an influential point (unusual x). Point #7 does not have an unusual x, in that it lies within the rest of the dataset (although near the high end). However, the residual of #7 is unusual in the dataset (large residual). In this case, Point #7 is an outlier (unusual y). We can also see that Cook’s Distance is large for both points #1 and #7 at ~0.55 and ~0.65, respectively. This indicates that both points change the parameter estimates by a significant amount when removed.

2B.

After refitting the linear regression to a modified dataset that excludes points #1 and #7, we see a much better fit of the model, which can be seen by the comparison of several values. The F value increased from 37.35 (old model) to 102.9 (new model). The RMSE decreased from 3,572.09 (old model) to 2,501.06 (new model). The R2 increased from 0.5545 (old model) to 0.7861 (new model). The t value of the least squares estimate for the “miles” coefficient decreased from -6.111 (old model) to -10.14 (new model). All of these changes are expected with modifying a dataset to remove outliers and influential points, which result in a more accurate and precise model: (1) the F value of the overall model should increase, signaling a stronger significance of the model; (2) the RMSE decreases, which shows that the standard deviation of the residuals decreases and that the new model is a better fit to the observed data points; (3) the R2 value increases, which signals that the model accounts for more variation; (4) the t value of the predictor parameter estimate increases, which shows that the new regression is a better fit to the dataset. This also agrees with Cook’s Distance for both points #1 and #7 that indicate that both points change parameter estimates by a significant amount when removed.

Question 3.

3A.

The new MLR model is definitely useful in predicting price of Prius cars. The R2 value for this MLR is 0.8031, which indicates that over 80% of the variance of the predicted price of Prius cars is explained by this model.

3B.

The four assumptions are still met with the new MLR model. The addition of a new predictor variable could potentially change those assumptions, as a new regression is fit to the dataset, which therefore produce new residuals and errors. In this particular example, we see that the assumptions are actually stronger, as there is a better mean linearity, homoscedasticity, and error normality.

3C.

To test the significance of the overall fit of the regression model, we will use a right-tailed F test with a null hypothesis (H0) of B1=B2=0, an alternative hypothesis (HA) of Bk≠0 for some k (at least one predictor not equal to 0). We are not given a significance level, but I will assume a significance level of α=0.05. If the F value of the model is beyond the F critical value, we will reject the null hypothesis. Otherwise we will fail to reject the null hypothesis.

Here see that the F value of the new MLR model is 59.15 and the F0.05,2,29 (just assuming a significance level of 0.05 for comparison here) critical value is 3.327654. The F value of the model is well beyond the critical value, and we reject the null hypothesis. This is also echoed by the p-value of 5.835e-11, which is much lower than common significance levels used. We can conclude that at least one predictor is not equal to zero and that the overall fit of the MLR model relating price of Prius cars (response) to mileage and age of vehicle (predictors) is significant.

3D.

We are not given a specific significance level criterion for assessing variable significance, but I will assume a significance level of alpha=0.05. Within this context, all predictor coefficients are significant as the p-values of age (1.39e-06) and mileage (0.0151) are both well below the chosen significance level.

This implies that mileage, while keeping age predictor constant, significantly changes the mean response of the predicted price of Prius cars. More specifically, we can be 95% confident that the price of a Prius car decreases by $0.02753 for each additional mile of use. Similarly, age, while keeping mileage predictor constant, significantly changes the mean response of the predicted price of Prius cars. More specifically, we can be 95% confident that the price of a Prius car decreases by $843.8 for each additional year of age (assuming year is age unit).

3E.

For reference, the original model has an R2 value of 0.5545 and an adjusted R2 value of 0.5397, and this new MLR model has an R2 value of 0.8031 and an adjusted R2 of 0.7895. The difference between explained variance is quite large, and I would choose the latter MLR to predict the sale price of Prius cars.

3F.

Both predictors are statistically significant in the new MLR. They are also practically significant. Although mileage is related to age in some sense, they are independent as style and amount of driving is unique to each driver.

Question 4.

4A.

Using these variables, one could examine if total spending and thus profit of Kroger (response variable) can be predicted by the variables number of items purchased, day of week of transaction, time of day of transaction, and payment type.

4B.

An additional explanatory variable that could be used would whether the transaction took place on a weekend or not. This would be categorical, and would be in the form of an indicator variable in the MLR.

4C.

The full regression equation using these parameters would be:

Total amount spent = Number of items purchased + Day of week + Time of day + Payment type + Weekend