COMPARING STUDY DESIGNS

1.
   a.
      i. Block design (age is the blocking factor)
      ii. Completely randomized design
      iii. Matched-pairs design (since subjects are matched together based on age and sex)

   b. Things to consider when answering this question:
      - All studies use random assignment, so all have the most important element of study design!
         - We don't have to worry about bias due to lurking variables in any of the studies
      - Studies (i) and (iii) have an advantage over study (ii) because they will have less variability to deal with when trying to estimate the treatment effect
         - Study (i) reduced variability in the response due to age by blocking on age
         - Study (iii) reduced variability in the response due to age and by matching on these variables
         - Meaning: we've already accounted for age (and sex) in the design, so the effect of age (or sex) on the response will not interfere with our ability to measure the effect of treatment
      - Problems such as non-adherence or bias due to the subjects or the researcher (if not blinded*) will likely impact all three studies, so we cannot consider these to help us determine which study is better
      - Considering all of this, I would choose study (i) or (ii) depending on how important I though the effect of biological sex is on the response

      * We were not told that any of the studies were blinded, so you should not assume they were when answering this question; in real life, however, there is very little point to giving a placebo if you are *not* going to blind at least the subjects, so all three studies were likely at least single blind.

2.
   a.
      i. Matched-pairs design (since each subject is measured with clicker and without)
      ii. Block design (section is the blocking factor)
      iii. Observational study (since students determined for themselves if they wanted to use clickers or not)

   b. Things to consider when answering this question:
      - Study (ii) uses random assignment, so we don't have to worry about bias due to lurking variables with this study
      - Study (i) has each student serve as their own control, which also avoids bias due to lurking variables (since each subject has the same lurking variables as themselves)
         - There could be a problem caused by the fact that all students when from no clickers to using clickers…for example, if we observe no improvement in performance, it may be because the material in the course got harder, but we might think it was because clickers are ineffective
      - Study (iii) is vulnerable to bias due to lurking variables, so this is the worst option of the three!
      - Again, problems such as non-adherence or bias due to the subjects or the researcher will likely impact all three studies, so we cannot consider these to help us determine which study is better

- Considering all of this, I would choose study (ii)

HYPOTHESIS TESTING: ALPHA, BETA, POWER

3.
    a. Saying that less than 30% of orders are shipped late when in fact the true proportion is 0.3.
    b. Several answers are possible. For example: The manufacture may lose customers who are frustrated with their late shipping and thus loose business/money.
    c. 6% of hypothesis tests with the same properties (e.g. hypothesis, rejection criteria, based on random samples of 50 orders) will incorrectly reject the null hypothesis.
    d. Saying that 30% of orders are shipped late when in fact it is less than 30%.
    e. Several answers are possible. For example: The manufacturer may make unnecessary changes to their production and/or shipping process. This will cost them money they didn't need to spend.
    f. 79% of hypothesis tests with the same properties will fail to reject the null hypothesis when it false and 21% will reject the null when it is false.
    g. 16% of hypothesis tests with the same properties will incorrectly fail to reject the null hypothesis, while 84% will correctly reject it.
    h. β is lower and power is higher for part (i) because we are assuming there is a larger effect; it is easier to distinguish between 0.3 and 0.15 than it is to distinguish between 0.3 and 0.25. This leads to higher power and thus lower probability of making a Type II Error.

4.
    a.

The POWER Procedure
One-Sample t Test for Mean

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Number of Sides	1
Null Mean	50
Alpha	0.05

Computed N Total					
Index	Mean	Std Dev	Nominal Power	Actual Power	N Total
1	55	20	0.8	0.802	101
2	55	20	0.9	0.901	139
3	55	30	0.8	0.800	224
4	55	30	0.9	0.900	310
5	60	20	0.8	0.812	27
6	60	20	0.9	0.903	36
7	60	30	0.8	0.806	58
8	60	30	0.9	0.902	79

b. As we require higher power, n increases (this is consistent with what we learned about increasing sample size to increase power). As the effect size ($\mu_{alternative} - \mu_{null}$) increases, n decreases (this is consistent with what we learned about power increasing if there is a bigger effect to find…so we could use a smaller sample size to achieve the same power). As the standard deviation increases, n increases (this is consistent with what we learned about power increasing if there is less variability).

c.

The POWER Procedure
One-Sample t Test for Mean

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Number of Sides	1
Null Mean	50
Alpha	0.05
Total Sample Size	75

Computed Power			
Index	Mean	Std Dev	Power
1	55	20	0.692
2	55	30	0.415
3	60	20	0.996
4	60	30	0.888

(You were not asked to comment on this output, but note that the predicted power is highest for the largest effect size with the smallest standard deviation and power is lowest for the smallest effect size with the largest standard deviation. This is consistent with everything we've learned about power [which was discussed in part (b) of this question].)

CHI-SQUARE TESTS

5.

a. Independence
b. Goodness of fit (Note: if we assume that Machines B and C produce defective parts as the same rate, then the null hypothesis would be $H_0: p_A = 0.50, p_B = 0.25, p_C = 0.25$.)
c. Homogeneity
d. Goodness of fit ($H_0: p_1 = p_2 = \cdots = p_8 = 1/8$)
e. Goodness of fit ($H_0: p_1 = 0.56, p_2 = 0.19, p_3 = 0.19, p_4 = 0.06$)
f. Independence
g. Homogeneity

ANOVA / REGRESSION ANALYSIS

6.

a. AAA: *SSR* = (87769891)(4) = 351079564
BBB: *SST* = 351079564 + 95176801 = 446256365
CCC: *MSE* = 95176801 / 65 = 1464258
DDD: *F* = 87769891 / 1464258 = 59.9415

b. $\hat{y} = -2774686 - (0.01602)(58000) + (1386.00789)(2012) + (12.88949)(271) + (2251.11817)(1)$
c. $R^2 = \frac{SSR}{SST} = \frac{351079565}{446256366} = 0.7867 = 1 - \frac{SSE}{SST} = 1 - \frac{95176801}{446256366}$
d. $s = \sqrt{MSE} = \sqrt{1464258} = 1210.065$
e. We would expect price to decrease by \$0.016, on average for each additional mile on the odometer, holding the other variables fixed. (Notice that the question is asking about the interpretation of the slope coefficient for mileage; you just need to recognize that and find the appropriate value in the output.)
f. Yes, according to the F-test for the overall model:
$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \; vs. H_a$: at least one $\beta_i \neq 0$
Test stat: F = 59.94
p-value < 0.0001 under the F-distribution with 4 and 65 degrees of freedom
This is less than 0.05, so we reject H_0 and conclude there is evidence that something in the model is useful for predicting price.
g. Yes, according to the t-test for the slope of horsepower:
$H_0: \beta_3 = 0 \; vs. H_a: \beta_3 \neq 0$
Test stat: t = 3.16
p-value = 0.0024 under the t-distribution with 65 degrees of freedom
This is less than 0.05, so we reject H_0 and conclude there is evidence that horsepower is useful for predicting price (after accounting for the other variables in the model).
h. Plot 1 is checking linearity and constant variance (it can also be used to check for outliers and if the error terms average to zero, but these were not asked about in the question). There may be concerns about the linearity condition, as there appears to be a non-linear pattern in the residual plot. (Note: it is either this or there are several outliers. To address the concern, start by adding a polynomial term to the model [you would need to look at the individual residual plots that are not shown here to determine which variable needs the polynomial] or try transforming the response. If that does not fix the problem, try identifying the possible outliers.) Plot 2 is checking normality. There are concerns about the error terms being normally distributed because the QQ-plot shows a curve rather that straight upward sloping line; however we don't really need to worry about this because the sample size (70) is larger than 30.

7.
a. $\hat{y} = 20633 - (0.06928)(58000) = \$16{,}614.76$
b. $16614.76 \pm (1.995)(2385.07332)\sqrt{\frac{1}{70} + \frac{(58000-31314.29)^2}{12383085714}}$
c. We are 95% confident that the true average price for all cars that have 58,000 miles on them is between \$15,340 and \$17,890.
d. If we were to take all possible samples of 70 cars from this population and calculate a 95% confidence interval for each, then 95% of those intervals would contain the true average price of all cars that have 58,000 miles.
e. $16614.76 \pm (1.995)(2385.07332)\sqrt{1 + \frac{1}{70} + \frac{(58000-31314.29)^2}{12383085714}}$
f. We are 95% confident that an individual value for any car that has 58,000 miles on it is between \$11,689 and \$21,541. (Note: a prediction interval is still a confidence interval at its heart, so it is still appropriate to say "We are 95% confident…")