Question 1. [12 marks]

(a) Should you do the test assuming paired or independent samples? Explain briefly.

[1] The samples are paired since a single sample of neighbourhoods was specified.

(b) Should your test be a parametric or a non-parametric test? Explain briefly with specific reference to the appropriate boxplot(s).

[2] For a paired test, we only examine the boxplot of **differences.** This graph is symmetric and has no outliers. Since it is reasonable to assume a normal distribution for the population of differences, a parametric test is appropriate.

For the independent samples test, examine the boxplots of the two samples. Again parametric test.

(c) Perform the appropriate test at the 5% level of significance. Use the critical value approach.

[4] `Paired T for P_highsch - P_collgrad`

```
               N      Mean     StDev   SE Mean
Difference    20  0.088571  0.053563  0.011977

95% lower bound for mean difference: 0.067861
T-Test of mean difference = 0.06 (vs > 0.06): T-Value = 2.39  P-Value = 0.014
```

-Ho: $\mu_1 - \mu_2 = 0.06$; Ha: $\mu_1 - \mu_2 > 0.06$
-t = (0.0886-0.06)/0.01198 = 2.39
-rejection region t > 1.73 (19 d.f.)
-reject Ho, conclude proportion of HS grads higher than proportion college grads by > 6%.

(d) If you performed a paired test in (c), then perform the independent samples test now. If you performed an independent samples test in (c), now perform a paired test.

[4] `Two-sample T for P_highsch vs P_collgrad`

```
               N    Mean   StDev  SE Mean
P_highsch     20  0.2660  0.0489    0.011
P_collgrad    20  0.1774  0.0595    0.013

95% lower bound for difference:  0.059508
T-Test of difference = 0.06 (vs >): T-Value = 1.66  P-Value = 0.053  DF = 36
```

- Ho: $\mu_1 - \mu_2 = 0.06$; Ha: $\mu_1 - \mu_2 > 0.06$
- t = (0.0886 – 0.06)/sqrt(0.0489^2/20 + 0.0595^2/20) = 0.0286/0.0172 = 1.66
-rejection region t > 1.69 (36 d.f.)
-do not reject Ho, cannot conclude the proportion HS grad higher by more than 6%.

If equal population variances assumed, pooled stdev is 0.0544 with 38 d.f. but no change in t-statistic.

(e) Why do the two tests in parts (c) and (d) result in different conclusions?

[1] Variation between neighbourhoods masks the difference between the two proportions. (The paired test accounts for the paired nature of the data and reduces the unexplained variation.)

Question 2. [13 marks]

(a) Test whether the proportions of neighbourhoods with median incomes over $25,000 are different in the two populations. Calculate a z-statistic and complete the test using the p-value approach and a 1% level of significance.

[4]

```
p_univgrad>0.25   X   N  Sample p
0                27  51  0.529412
1                49  59  0.830508

Difference = p (0) - p (1)
Estimate for difference:  -0.301097
99% CI for difference:  (-0.520736, -0.0814579)
Test for difference = 0 (vs not = 0):  Z = -3.41  P-Value = 0.001
```

-Ho: p1-p2=0; Ha: p1-p2 ≠ 0
-pooled p = 76/110 = 0.69 *(0.5 mark)*
-z = 0.30 / sqrt(0.69*0.31*(1/51+1/59)) = 0.30/0.0884 = 3.4
-p-value is 2*0.0003 = 0.0006 *(0.5 mark)* *same deduction for using critical value approach*
-since p-value < 0.01, we reject Ho, conclude there is a difference
(1 mark for each point above, except where noted)

(b) Calculate a 99% confidence interval to estimate the difference in the two population proportions. Explain how this interval leads to the same conclusion as above.

[2] 0.30 ± 2.58 * sqrt(0.53*0.47/51 + 0.83*0.17/59)

= 0.30 ± 2.58 * 0.085 = 0.30 ± 0.22 = (0.08, 0.52)

Deduction of 0.5 mark for incorrect difference, 0.5 for incorrect z-critical value, 0.5 for incorrect standard error, 0.5 for not noting interval excludes the hypothesized value 0.

(c) Fill in the blanks: The test and confidence interval above assume that the <u>sample proportions</u> have a <u>normal</u> distribution. We are assured this is true because <u>the sample sizes are large enough to apply CLT (or x ≥ 10 and n-x ≥ 10 in each sample).</u>

[3] *These answers go together—I gave no marks for "data or populations have a normal distribution".*

(d) Now do the same test using a chi-square test of homogeneity. Use the p-value approach and a 1% level of significance.

[4]

```
             Total
    1     27     24     51
       35.24  15.76
       1.925  4.303

    2     49     10     59
       40.76  18.24
       1.664  3.720

Total     76     34    110

Chi-Sq = 11.613, DF = 1, P-Value = 0.001
```

-Ho: p1=p2; Ha: p1 ≠ p2
-χ^2 = 11.61
-p-value < 0.001 (from table) which is < 0.01
-reject Ho, conclude proportions are different

Acceptable-- Ho: Income level and Education (HS vs college) are independent.
Not acceptable—Ho: Factors A and B are indep.

Question 3. [8 marks]

On April 19, 2016, Manitoba held its provincial election. The morning after the vote, the breakdown of the vote was: Progressive Conservative 53.7%, NDP 25.2%, Liberal 14.4%, other 6.7%.

As of April 12, the latest poll showed a sample breakdown of 51.3% PCs, 25.9% NDP, 18.2% Liberal, and 4.5% other.

(a) Assuming a sample size of 1000 for the vote projection, test whether the projection could be considered an accurate reflection of the distribution of the actual vote one week later. Use a 5% level of significance and the critical value approach.

[4]

Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: n

Category	Observed	Test Proportion	Expected	Contribution to Chi-Sq
1	513	0.537	537	1.0726
2	259	0.252	252	0.1944
3	182	0.144	144	10.0278
4	46	0.067	67	6.5821

N	DF	Chi-Sq	P-Value
1000	3	17.8769	0.000

Ho: p1=.537,p2=.252, p3=.144, p4=.067; Ha: one pi not as stated
-Chi-sq stat is 17.88
-Rejection region is > 7.81
-Reject Ho, conclude projection not accurate.

1 mark for each point above. I deducted 2 marks if they treated the electoral results as the sample data—the confusion was that the idea that the poll projections became the results that people "expected".

(b) Now suppose that in a sample of 25 university students in St. Boniface, Manitoba, only one voted for the Liberal party. Test whether this constitutes sufficient evidence to show that the proportion of university students in St. Boniface who voted for the Liberal party is less than 14.4%.

[4]

-Ho: p = 0.144; Ha: p < 0.144
-Cannot use normal approx. since np = 25*0.144 = 3.6 < 10
-p-value is P(X ≤ 1) = 0.106 not < 0.05, based on n=25, p=0.144
-Cannot reject Ho, conclude insufficient evidence to show p < 0.144.

p-value is P(0) + P(1) = 0.856^25 + 25*0.144*0.856^24 = 0.0205 + 0.0862 = 0.106

1 mark for each point above

*If the solution calculates z = (0.04 – 0.144)/sqrt(0.144*0.856/25) = -0.104/0.07 = -1.49, with a rejection region z < -1.645, then give only 2 marks for hypotheses and decision/conclusion.*

Question 4. [18 marks]

(a) Here is part of the analysis given as an output table. Complete the missing entries in this table. Show your calculations for only MSE, R-Sq, and 's'.

Analysis of Variance for ProcessTime

Source	DF	SS	MS
Shift	2	159.849	79.9244
Machine	2	7.502	3.7511
Machine*Shift	4	97.662	24.4156
Error	18	119.547	6.6415
Total	26	384.560	

S = 2.577 R-Sq = 68.91% R-Sq(adj) = 55.10%

[4]
$$R^2 = 1 - \frac{SSE}{SSTot} = 1 - \frac{119.547}{384.560} = 1 - 0.3109 = 0.6891$$
$$MSE = \frac{SSE}{df_E} = \frac{119.547}{18} = 6.6415$$
$$\{s = s_e = s_P\} = \sqrt{MSE} = \sqrt{6.6415} = 2.5771$$

Deduct 0.5 mark for each incorrect answer above. If there is a loss of 0.5 mark for the wrong df, the incorrect MS based on the incorrect df should not be docked another 0.5 mark. However, since the 18 df for the SSE is implied by the t-values in part (f), there should be a full mark deduction if df ≠ 18 and MSE is also wrong, but do not also deduct for the wrong s if it is the sqrt of the wrong MSE value.

(b) Now show an alternative method of calculating the MSE which combines the standard deviations in the appendix (do not complete the calculation, just show how they would be combined).

By Pooling the treatment variances as follows:

$$MSE = s_p^2 = \frac{\sum_{i,j=1}^{i,j=3} (r-1)s_{ij}^2}{n-ab} = \frac{\sum_{i,j=1}^{i,j=3} (r-1)s_{ij}^2}{ab(r-1)}$$ or simplified as $(2*3.143^2 + \ldots + 2*0.231^2)/(2*9)$

[1]
$$MSE = s_p^2 = \frac{\sum_{i,j=1}^{i,j=3} s_{ij}^2}{ab} = \frac{3.143^2 + \ldots + 0.231^2}{(3*3)} = \frac{59.7712}{9} = 6.6412$$

(a) What do you observe about the possible effects of Shift and Machine (separately and jointly) on the average processing time? Justify your answer with clear references to the relevant plot. Which machine results in the same average processing time, regardless of the shift?

[3] -Joint Effect: From both the 'Interaction' plots, it can be seen that the plot lines are not parallel; in one plot, they actually intersect and in the other they converge on a point. This is clear indication of the fact that 'Interaction' may be present.

-Separate effects: From the interaction plot of means against machines, the night shift appears to be less efficient. It is less clear whether there is a maineffect due to the machine.
-The effect of Machine 1 does not seem to change with the shift, unlike the other two machines.
1 mark for each point above

(d) Test whether the effect of the Shift on the processing time depends on the Machine. Use a 5% level of significance.

[3] S1- H_0: Interaction is not Present; H_a: Interaction IS Present.

S2- F_{Calc} =(MSInter)/MSE = (24.4156)/6.6415 = 3.6762

S3- Rejection region is $F > F_{Crit} = F_{0.05}(df_{Int} = 4, df_E = 18) \approx 2.946$

S4- Since $\{ F_{Calc} = 3.6762\} > \{ F_{Crit} = 2.946\}$ ==> Reject H0

S5- Based on available statistical evidence, one can state that the effect of 'Shift' on processing time, does depend upon the 'Machine', or simply, there is interaction between Shift and Machine

-1 mark for hypotheses, 0.5 mark for correct F-stat, 0.5 mark for rejection region, 0.5 for decision, 0.5 for conclusion

(e) If one were to do a test to compare the mean processing times for the three shifts and the p-value was found to be 0.000, what would you conclude? Given the result in part (d) above, would you consider this conclusion meaningful? Explain why you would consider it meaningful or not.

(H_0: $\mu_1 = \mu_2 = \mu_3$ H_a: Not all $\mu_{i's}$ are equal.

Since {p_Val = 0.000} < {$\alpha = 0.05$} ==> Reject H0)

[2] -Based on the p-value, one could claim that the mean process times by machine are not equal.
-However, since 'Interaction' is present, this difference may not be meaningful since there is no Shift effect with Machine 1 but probably some with the other machines.
An alternative view: If the company wanted to cut back on the shifts but run all three machines, it might consider cutting out the night shift since it is the least efficient shift. In this sense, one could say that the main effect of the Shift factor is meaningful.
1 for conclusion from the test for main effect, 1 for a reasonable comment on whether it is meaningful.

(f) Using the Bonferroni method of multiple comparisons, calculate the margin of error for comparing the treatment means. What do you conclude about the mean process time in minutes for machine 3 operated in the morning and machine 2 operated in the night? Which is the best combination?

{Hint: You will need one of the following values for $t^* = t_{\alpha/(2m)}(df_E)$ where 'm' is the number of pair-wise comparisons. Based on the correct value of 'm', **choose the appropriate** value given here:
$t_{0.025} = 2.1009$, $t_{0.00833} = 2.6393$, $t_{0.00625} = 2.7745$, $t_{0.002778} = 3.1482$, $t_{0.001389} = 3.4622$, or $t_{0.000694} = 3.7745$}

[5] Here 'r' = 3, a = 3, b = 3 and sp = 2.5771
Thus m = ${}_{ab}C_2 = {}_9C_2 = (9!)/\{(9-2)!*2!\} = 9*8/2 = 36$

$t^* = t_{\alpha/2m}(df_E) = t_{0.05/(2*36)}(18) = t_{0.000694}(18) = 3.7745$

$$ME = t^* s_p \sqrt{\frac{2}{r}} = 3.7745 * 2.5771 * \sqrt{\frac{2}{3}} = 7.9423$$

$$| \bar{X}_{23} - \bar{X}_{32} | = | 35.27 - 45.80 | = 10.53$$

Since $| \bar{X}_{23} - \bar{X}_{32} | = 10.53\} > \{ME = 7.9423\}$ ==> There is a difference in the processing time between machine 3 operated in the morning and machine 2 operated in the night.

[5]

The best combination is when the 'Treatment Mean' or the 'ProcessTime' is the smallest. This happens when machine 3 is operated in the morning. However, given the ME of 7.94, we can only say with some certainty that this combination is better than either Machine 2 or 3, operated at night, and not necessarily the best compared to the other combinations.

-1 mark for choosing t = 3.7745*
-1 mark for the standard error of 2.5771 √(2/3) = 2.1*
-0.5 mark for finding a difference of 10.5 and 0.5 for noting it is > ME and
-1 mark for concluding there is a real difference
-1 mark for some reasonable comment on the best (better) combination

Question 5. [24 marks]

(a) Examine the residual plot for Model 1. Discuss any problems with the validity of the linear model assumptions.

[2] -many outliers beyond 3 standard errors, suggesting problems with assumption of normally distributed errors;

-a pattern of increasing variance, suggesting problems with constant error variance assumption.

-1 mark for the two assumptions; 1 mark for observations on residual plots.

(b) Now examine the residual plot for Model 2 (the dependent variable here is the logarithm (base 10) of price). Have the problems above been resolved?

[2] the residual plot has resolved the problem of non-constant variance, but there are still some extreme residual values beyond 3 standard errors, suggesting a problem with the assumption of normally distributed errors.

(c) Model 3 drops a number of observations from Model 2. What changes do you see from Model 2 to Model 3?

[2] Model 2:

```
The regression equation is
log10Price = - 1.19 + 0.000145 SqFt - 0.00120 Bedrms + 0.0149 Bathrms
      + 0.0112 AC + 0.0175 GarageNoCars + 0.0218 Pool + 0.00168 Year
      - 0.0658 Quality - 0.00747 Style + 0.000002 LotSize
      - 0.0389 Highway
S = 0.0778770   R-Sq = 83.1%   R-Sq(adj) = 82.7%
```

Model 3:

```
The regression equation is
log10Price* = - 1.79 + 0.000155 SqFt + 0.00071 Bedrms + 0.0206 Bathrms
       + 0.00696 AC + 0.0142 GarageNoCars + 0.0361 Pool + 0.00195 Year
       - 0.0502 Quality - 0.00855 Style + 0.000002 LotSize
       - 0.0374 Highway
S = 0.0711666   R-Sq = 85.6%   R-Sq(adj) = 85.3%
```

Some of the coefficients change in value, some more than others (generally within one std error)
As expected, the fit statistics are better.

(d) Test at the 5% level of significance whether Model 3 is useful for prediction.

[3]

Ho: beta1=0, …, beta11=0; Ha: at least one betaj nonzero
F =270.12 with p-value of zero
Reject Ho, conclude model is useful for prediction

(e) Test at the 1% level of significance whether the SqFt variable is an important variable in Model 3. Include the hypotheses, test statistic, rejection region, decision, and conclusion.

[4]

Ho: beta(SqFt) = 0; Ha: beta(SqFt) nonzero
t = 0.000155 / 0.00000887 = 17.5,
since |t| > 2.58, we reject Ho,
conclude SqFt variable is useful, given the other variables in the model
1 mark for each point above, with 0.5 mark for the “given…”

(f) What does the regression coefficient of the SqFt variable estimate? Be as precise as possible and use the appropriate units of measurement.

[2]

If houses increase in size by one square foot,
there is an (estimated) increase of 0.000155
in the logarithm of price (on average) (price in thousands of dollars),
assuming all other characteristics remain unchanged.
0.5 mark for each point above (but do not insist on the content in brackets)
(note that this is a multiplicative increase of 10^0.000155 = 1.000357 in the price in dollars,
That is, a $100,000 house increases to $100036, a $500,000 house to $500,178, etc.

(g) Explain whether multicollinearity is an issue that might affect the regression coefficients.

[1]The Variance Inflation Factors (VIF) for all variables are small < 5, indicating there is no problem with multicollinearity; therefore, the coefficients are not affected.

(h) What is the interpretation of the coefficient of the Pool variable?

[2]

Houses with a pool compared to those without a pool,
have an (estimated) increase of 0.036
in the logarithm of price (on average)
assuming other variables remain unchanged.
(0.5 mark for each point above, but you may ignore the brackets).

(This is a multiplicative increase of 10^(0.036) = 1.0864 in the price in dollars)

(i) Model 4 calculated 99% intervals for the logarithm (base 10) of the price of a house with the characteristics as shown under "Value of predictors", but these intervals have been erased. Calculate the appropriate 99% interval for the price (in logarithmic terms) of a house with the given characteristics.

[3]

Fit = 2.91730, SE Fit = 0.01928.
SE(prediction) = sqrt(0.0051 + 0.01928^2) = 0.074
99% prediction interval is 2.9173 ± 2.58 * 0.074 = 2.9173 ± 0.1908 = (2.7265, 3.1081)

(j) What is the estimated mean price of houses with the same characteristics? Use the appropriate units of measurement. (If you cannot complete the calculation, then show how it would be calculated and give an approximate value.)

[1]

10^2.9173 = 826.609 thousands of dollars,
Note that ID 72 has a log10price of 2.919 and a price **$830,000** (close enough)

Showing **10^2.9173 ($thousands)** or using 10^3 = $1000 thousands is worth 0.5 mark, if the proper units are shown. For the full mark, either $826,000 or $830 thousands is necessary, with a 0.5 deduction if the proper units are not shown.

(k) According to the output from the Best Subsets procedure, which model might be considered the best? Answer using two different criteria.

[2]

The 9 variable model that drops only the Bedrms and the AC variables.
It has the lowest s of 0.071066 (and the highest adj R-sq of 85.2) and the lowest Cp of 8.6.
Using the lowest s and highest adj R-sq are the same criterion, and worth only 1 mark.