STAT 6338
Advanced Statistical Methods II
Project 1

Name: Harshul Shah
NetID: hxs230024

Ans 1)
a. Hypotheses:
 Null Hypothesis (H0): The mean infection risk is the same across all four regions.
 Alternative Hypothesis (Ha): The mean infection risk is not the same across all four regions.
 ANOVA Results:
 F-value: 3.21 Pr > F: 0.0201
Conclusion: Since the p-value (0.0201) is less than the significance level of $\alpha = 0.05$, we reject the null hypothesis. We conclude that there is a statistically significant difference in the mean infection risk among the four geographic regions.

b. Pairwise comparisons are between each combination of means.
 Regions 1 vs 2: Lower CL diff
 Regions 1 vs 3: Lower CL diff
 Regions 1 vs 4: Lower CL diff
 Regions 2 vs 3: Lower CL diff
 Regions 2 vs 4: Lower CL diff
 Regions 3 vs 4: Lower CL diff
 For Tukey's test to reveal a difference, the "diff" column (i.e. differences in means) must not cross 0. If they cross 0 or 0 is in between Upper CL and Lower CL, the analysis cannot show anything.

Ans 2)
Effect of average age of patient (variable 3) on mean infection risk (variable 4) is to be studied. For purposes of this ANOVA study, average age is to be classified into four categories: under 50, 50-54.9, 55.0-59.9, 60.0 and over. Assume that ANOVA model is applicable.
Test whether or not the mean infection risk differs for the four age groups. Control the α risk at .10.
We categorized average patient age into four groups and performed ANOVA to determine if there were significant differences in mean infection risk among these groups. The results from the PDF show the following:
F Value: 0.56
Pr > F: 0.6412
Conclusion:
Since the p-value is not below 0.10, we fail to reject the null hypothesis and conclude that there is no significant difference in mean infection risk across the four age groups.

Ans 3)
b. Brown-Forsythe Test Results:
F Value: 4.33
Pr > F: 0.0064
Explanation: The Brown-Forsythe test assesses whether the variances in length of stay are equal across the four geographic regions. A significant p-value suggests unequal variances.

Conclusion:
The Brown-Forsythe test's p-value is 0.0064, which is less than the conventional alpha level of 0.05. Therefore, we reject the null hypothesis and conclude that the variances in length of stay are not equal across the regions.

c. Mean Length of Stay by Region:
Region 1: 11.0889
Region 2: 9.6834
Region 3: 9.1914
Region 4: 8.1138
Standard Deviation of Length of Stay by Region:
Region 1: 2.6696
Region 2: 1.1929
Region 3: 1.2250
Region 4: 1.0031
Explanation: Calculated the means (Yi) and standard deviations (si) for length of stay in each region to determine the appropriate transformation. Comparing Yi and si the si increases as Yi increases, we would consider transformations (logarithmic) to stabilise the variance.

e. Transformed Data ANOVA Results:
F Value: 14.79
Pr > F: <.0001
Explanation: Fitted an ANOVA model to the reciprocal transformed length of stay data to assess differences across geographic regions. The very low p-value indicates significant differences.
Conclusion:
The F is 14.79 and the p < 0.0001 which is highly significant, we reject the null hypothesis and conclude that the mean reciprocal transformed length of stay differs significantly across the four regions.