A. Given that the data are normally distributed from a random sample, we use the appropriate function to calculate the confidence interval by using the critical value of t distribution as the standard deviation of the population is unknown. In each interval the results were the same as the interval that the t.test command provides. We get the confidence interval of (31.8595,35.00412) at 95% for the average salaries of men.

B. Similarly, the confidence interval for the average salaries of women is (31.60332,34.79668).

C. Similarly, the confidence interval for the mean difference of the average salaries between the two genders is (-0.07852,0.54216).

D. H0: average salary of men =average salary of women

H1: average salary of men >average salary of women

At a significance level of a=0.05, from command t.test we obtain a p-value=0.1063. a<p-value Hence, we cannot reject H0, that the average salary of men equals average salary of women.

E. H0: True ratio of variances is equal to 1

H1: true ratio of variances is not equal to 1.

For the hypothesis of the equality of variances, the var.test is used at confidence level of 0.99. Since p-value=0.9445 and a=0.01 a<p-value. Hence, we cannot reject Ho, that the variances of the average salaries of the different genders are equal.

F. H0: True mean is equal to zero

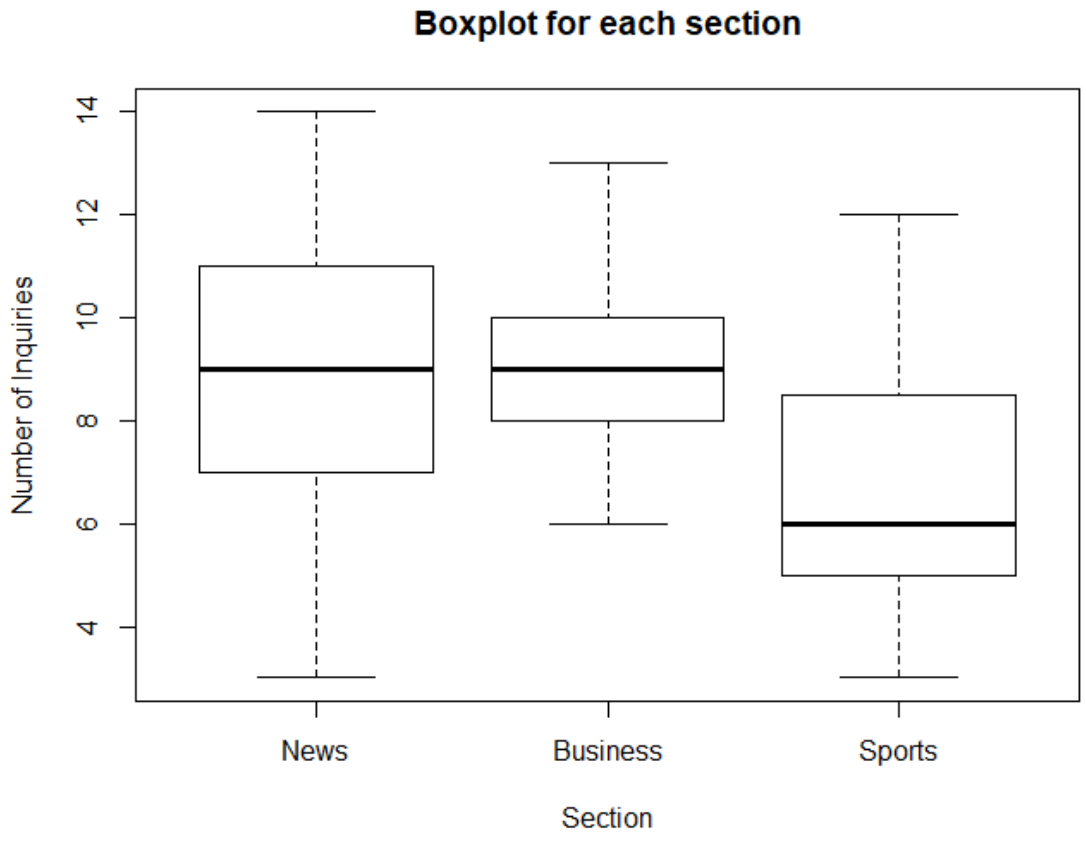
H1: True mean is not equal to zero

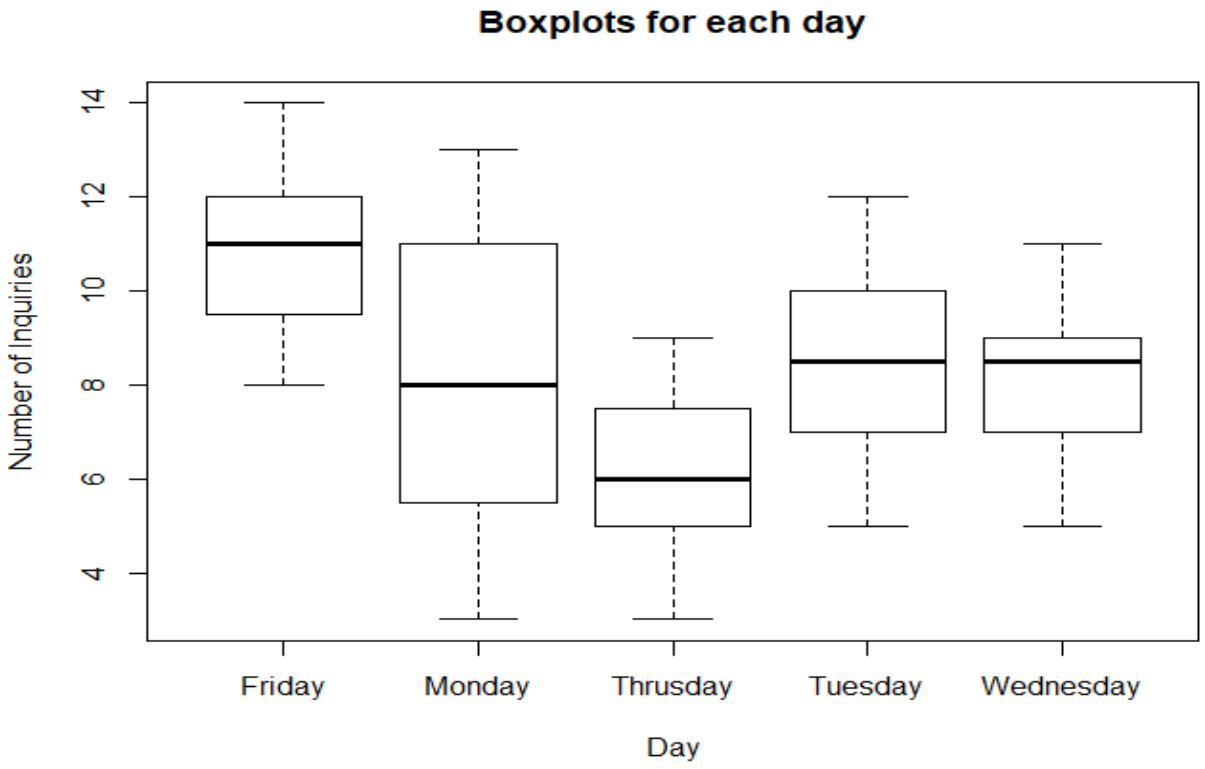
For this test, the t.test command was used with a confidence level of 0.9 and equal variances. P-value=0.2127 and a=0.1, so a<p-value. Hence, we cannot reject H0 that the average salaries of each gender are equal. There is not evidence for the statistical significance of gender on average salaries.

Exercise 2

First, the data frame “inquiries” is created by adding all the number of inquiries in column “val”, the name of each day in column “day” and the name of each section in column “section”. Section=1 corresponds to section “News”, section=2 to “Business” and section=3 to “Sports”.

A. In the boxplot for each section, we observe that News and Business section seem to have similar mean. Therefore, the statistical significance of using both may not be high. On the contrary, the section Sports seems to have a different sample mean than the other two. Regarding boxplots for each day, days like Tuesday and Wednesday seem to have similar means, while Friday and Thursday seem to have different means.





B. Given that the data are normally distributed, we have to test the assumptions for homogeneity of variances of the residuals before fitting the ANOVA model. By using the Bartlett test, we get a p-value of 0.098. Since a=0.05<p-value we cannot reject H0 that the variances of residuals in each category are equal. The same results are obtained by similar tests like Flinger test and Levene test. Next, the ANOVA model is used, and the results point that the p-value is really small and statistically significant at 0.01%. Hence, we reject H0 that all the means of each category are equal. There is at least one mean that is different from the rest. The model parameters are all significant. The intercept means that the average number of inquiries on Friday is 10.916 and all other days that have a negative coefficient mean that that they have a smaller number of inquiries in this day (for example Monday has 2.83 less inquiries than Friday in average).

C. By using Tukey HSD, we get all the combinations of each day and their statistical significance. The expected difference of Tuesday and Wednesday has a p-value of 0.9959 meaning that we cannot reject H0 that these days have equal means (or that their difference is different from zero and has statistical significance).

D. Given that the data are normally distributed, we have to test the assumptions for homogeneity of variances of the residuals before fitting the ANOVA model. By using the Bartlett test, we get a p-value of 0.06069. Since a=0.05<p-value we cannot reject H0 that the variances of residuals in each category are equal. The same results are obtained by similar tests like Flinger test and Levene test. Next, the ANOVA model is used, and the results point that the p-value=0.0193 is really small and statistically significant at 5%. Hence, we reject H0 that all the means of each category are equal. There is at least one mean that is different from the rest. By using Tukey HSD, we get all the combinations of each day and their statistical significance. The expected difference of 1 and 2 (News and Business) has a p-value of 0.9658 meaning that we cannot reject H0 that these sections have equal means (or that their difference is different from zero and has statistical significance). The most significant difference is between 3 and 2 (sports and business). The model parameters are not all significant. The section 2 (Business section) is not significant. The intercept is section 1 (News) and has a coefficient of 8.9 that means that the average number of inquiries for section 1 is 8.9. The coefficient for section 3 (Sports) means that the number of inquiries is 1.9 less on average than section 1.

E. Based on the previous findings we exclude section "News" and re-estimate the parameters. Given that the data are normally distributed, we have to test the assumptions for homogeneity of variances of the residuals before fitting the ANOVA model. By using the Bartlett test, we get a p-value of 0.09772. Since a=0.05<p-value we cannot reject H0 that the variances of residuals in each category are equal. The same results are obtained by similar tests like Flinger test and Levene test. Next, the ANOVA model is used, and the results point that the p-value=0.00451 is really small and statistically significant at 1%. Hence, we reject H0 that all the means of each category are equal. There is at least one mean that is different from the rest. By using Tukey HSD, the expected difference of the Business and Sports sections have a low p-value meaning that the do not have equal means. Now the parameters of the model are all significant. The intercept which is section 2 has a coefficient of 9.1 meaning that on average the number of inquiries is 9.1. The coefficient of section 3 is -2.1 meaning that it has 2.1 less number of inquiries on average than section 2.

F. We need to run the Anova model for the main effects model (without interaction) of factor day and section. We have already tested the assumptions for homogeneity of variances of the residuals of these two factors. By running the Anova model, we observe that both factor day and section have a significant and really small p-value meaning that we reject H0 that all the means of each category are equal. There is at least one mean that is different from the rest. By running Tukey HSD test we observe that several combinations are insignificant like Tuesday-Monday, Wednesday-Monday, Wednesday-Tuesday, and section 2-1. The parameters of the model are all significant except section 2 (Business). The intercept which is day Friday and section 1 has coefficient of 11.48 meaning that on Friday the section 1 has a number of inquiries on average of 11.48. The other significant parameters are all negative, meaning that for example a combination of Tuesday and section 3 have a small number of inquiries than the intercept by 2.41+1.9.

G. By using the stepwise method (command step or stepAIC) starting from the full main effects model, we observe that the AIC of the full main model is 90.51. By excluding each of the factors the AIC increases, therefore the model needs to have both factors as they are both significant. The significance of each coefficient was explained in the previous question.

H. Model 1 is the model with the full main effects, and we also create model 2 which is constant (8.33). Then, we run Anova between these two models. The results suggest that going from model 2(constant) to model 1 (full main effects) there are 6 more parameters added (df=6), the residual sum of squares is lower, and the p-value is significant at 0.1% meaning that the model with full main effects is different from constant model.