Project Part 3

Philip Lee

```
## Loading required package: carData
```

Question

The question that I want to answer is: is the average sales price of houses in the real estate market affected by the neighborhood where the houses are located at?

Data Description

(Reference 1 for source of the data) The data used in this project were collcted and were used to examine the relationship between the sale price of a real estate and three independent variables: appraised land value of the real estate, appraised value of improvements on the real estate, and the neighborhood in which the real estate is listed. The data consist of the appraised land values and improvement values and sale prices for real estates sold in city of Tampa, Florida, during the period between May 2008 and June 2009. The data pertains to eight neighborhoods (Hyde Park, Cheval, Hunter's Green, Davis Isles, Avila, Carrollwood, Tampa Palms, and Town & Country), which are relatively homogeneous but differ sociologically and in real estate types and values. The sales and appraisal data were recorded along with the neighborhood that the real estate was sold at. In total, the data has 350 observations. Each neighborhoods have the follwoing number of observations: AVILA - 12, CARROLLWOOD - 32,CHEVAL - 44, DAVISISLES - 42, HUNTERSGREEN - 56, HYDEPARK - 34, TAMPAPALMS - 75, TOWN&CNTRY - 55. Since the sample size for Avila neighborhood is too small, it will be excluded from the testing. For this part of the project, I will be focusing only on 2 out of 5 variables possible, which are: SALES variable, which is the sales price of the property in thousands of dollars, and NBHD variable, which is the neighborhood for which the property was sold at.

Data Relevance

The data I will be using is greatly related to the proposed question since the data has the sales prices of houses in different neighborhoods. This information can be used to determine the average sales prices of houses in different neighborhoods, which then can be used for hypothesis testing to answer the question by comparing the means of the sales prices in different neighborhoods.

Generalization

The null hypothesis for the test will be that the treatment means for the sales price will be equal to each other and the alternative hypothesis will be that at least one pair of the treatment means will be different. The result for the testing can be generalized to the actual population, in this case will be the real estate markets. If we reject the null hypothesis, then the outcome will be that the neighborhood that the house is located does affect the sales price of the house and vice versa.

Hypothesis Testing

The test that I will be conducting will be one-way ANOVA test. I wanted to use ANOVA test since I wanted to compare the sales price means for all the neighborhoods, which has more than 2 kinds. ANOVA has 3 assumptions that need to be satisfied: 1) the samples collected for each population are independent, 2) populations for each treatment are normal, 3) population variance for treatments are equal. For the test, I will consider each neighborhood as an individual population. Even though the neighborhoods come from the same city, they will have different demographics, classifications of the area (such as urban/suburban/rural), and many other factors that differentiate them from each other that they will be greatly different from each other to be considered as independent. The second assumption may also satisfied since ANOVA is robust with respect to the normality assumption as the sample sizes for each neighborhood is relatively large, which are greater than 20 (Reference 2). However, the normality assumption will be checked to be sure that it is not severely violated through Shapiro-Wilk test.

```
res.aov <- aov(SALES ~ NBHD, data = sale4)
aov_residuals <- residuals(object = res.aov )
shapiro.test(x = aov_residuals )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.73971, p-value < 2.2e-16
```

Since the p-value is less than the significance level of 0.05, there is a statistical evidence that the normality assumption is violated. The third assumption for variance needs to be tested through Levene's test at the significance level of 0.05.

```
leveneTest(SALES ~ NBHD, data = sale4)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   6  8.2587 2.407e-08 ***
##       331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since Levene's test of equal variances assumption has p-value less than the alpha, there is statistical evidence that the variance across the neighborhoods is statistically different. Due to the violations of normality and equal variances assumptions, the classic ANOVA test will not be used. Instead, Kruskal-Wallis rank sum test will be used as ANOVA assumptions are not met (Reference 3 for all of the codes used for the tests conducted).

```
kruskal.test(SALES ~ NBHD, data = sale4)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SALES by NBHD
## Kruskal-Wallis chi-squared = 190.11, df = 6, p-value < 2.2e-16
```

Since the p-value is less than the significance level of 0.05, we reject the null hypothesis and conclude that at least one pair of treatment means is different. Based on the tests, it can be concluded that there is a statistical evidence that the treatment means for each neighborhood are different from each other. Hence, it can be concluded that the neighborhood for which the house is located at does affect the average sales price of houses in the real estates market.

References

1. https://www.hcpafl.org/
2. Sinich, Terry. "The Analysis of Variance for Designed Experiments." A Second Course in Statistics Regression Analysis, by William M. Mendenhall, 7th ed., CRC Press, Taylor & Francis Group, 2016, pp. 693-693.
3. http://www.sthda.com/english/wiki/one-way-anova-test-in-r