Project Part 3

Research Questions

1. Was it normal that a property sold in Washington State in 2014 that had 4 bedrooms was much more expensive than one with three bedrooms?
2. Was it normal that a property sold in Washington State in 2014 with condition 4 out of 5 was much more expensive than one with condition 3? If not, was a property sold in Washington State in 2014 with condition 3 much more expensive than one with condition 4?

Data Description

The dataset, "House price prediction" was from Kaggle and was originally used to analyze and predict the future property prices in the real estate market. It was a sample that was collected among different selling properties in different cities in the state of Washington, USA. (Reference 1) The data contained 4600 properties with 18 features regarding this property, which were sale date, price, number of bedrooms, bethrooms, sqft living, sqft lot, floors, waterfront, number of view, condition, sqft above, sqft basement, year built, year renovated, street, city, statezip, and country.

Data Relevant

As the data contained a large sample of properties in the state of Washington in 2014, and bedrooms and conditions were two variables inside the data set. The research questions asked about prices of properties sold in Washington State in 2014 with certain features, which matched with what I had in my data set. The questions could be answered by checking if the mean of one level was significantly higher than another, and my data set should be a reasonable representation.

Both bedrooms with level 3 and 4 (three bedrooms and four bedrooms) and conditions 3 and 4 were within the variables' ranges in my data set, so there should be no extrapolation. As a result, this data set should be good to use to answer the research questions.

Generalization

People generally considered many factors before they bought properties, and numbers of bedrooms and condition ratings were two important factors. It was common that a property with three bedrooms met the needs, but a four-bedroom property might be better. At this time, people needed to consider that if it was worthy to pay a certain amount of money to buy a four-bedroom property. This was similar for condition ratings as well. This was why the outcomes of these two research questions could be important.

As the sample method of this data set was unclear, I did not know if this data set was a good representation of the whole population of properties sales in Washington (Reference 1). As a result, to have a relatively accurate and reliable conclusion, I should only generalize my outcomes to properties that fall into all the ranges of our independent variables. In conclusion, the outcomes of my tests could be reasonably generate to the properties sold in Washington in 2014, and it could be more accurate and reliable if generated to a population of properties that fall within all the ranges of my data set.

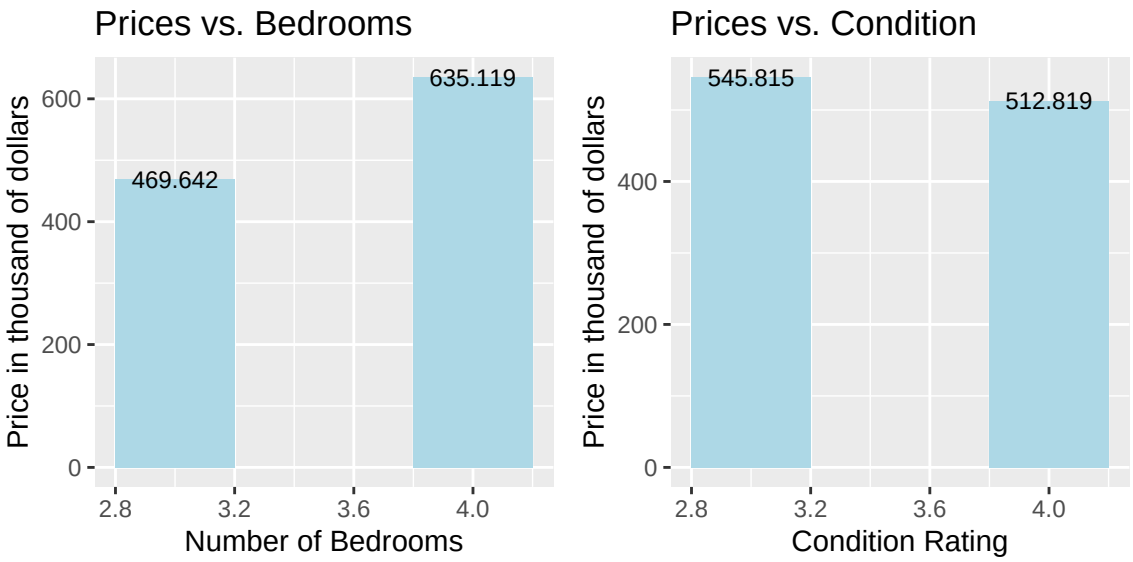
Data Analysis

Data Cleaning

Since there were two extreme outliers in this data set, I used filter to remove them. Also, to make the data easier to read, I made price unit from dollars to thousand of dollars.

Exploratory Data Analysis

```
bed <- data1[which(data1$bedrooms==3|data1$bedrooms==4),]
g1<-ggplot(bed, aes(x=bedrooms,y=price)) + stat_summary(fun.y=mean,
  geom="bar",fill = "lightblue", width =0.4)+labs(title="Prices vs. Bedrooms",
      x="Number of Bedrooms", y="Price in thousand of dollars",size =1) +
 stat_summary(fun.y=mean, geom="text", aes(label=round(..y..,3)),size=3)
condit <- data1[which(data1$condition==3|data1$condition==4),]
g2<-ggplot(condit, aes(x=condition,y=price)) + stat_summary(fun.y=mean,
  geom="bar",fill = "lightblue", width =0.4)+labs(title="Prices vs. Condition",
       x="Condition Rating", y="Price in thousand of dollars",size =1) +
    stat_summary(fun.y=mean, geom="text", aes(label=round(..y..,3)),size=3)
grid.arrange(g1, g2, ncol = 2)
```



For the first research question, I could tell from the graph of "Price vs. Bedrooms", the mean price of properties with condition 4 was higher than that of properties with condition 3. However, I did not know if this difference was significant or not, which applied to the question that if the it was "normally much more expensive", unless we conduct a hypothesis test. As a result, I decided to use a two-sample t test with null hypothesis H0: mu(bedroom 3) = mu(bedroom4), and alternative hypothesis Ha: mu(bedroom 4) > mu(bedroom3), to check if this difference between means was significant.

For the second research question, I could see from the graph that, on average, the price of properties with condition 4 was lower than the price of properties with condition 3. For the second research question, I could conclude that on average, a property sold in Washington State in 2014 with condition 4 was not more expensive than one with condition 3. Based on the EDA, I want to test the second part of the second research question that if the price of properties with condition 4 was significantly lower than that with condition 3. As a result, I decided to conduct a two-sample t test with null hypothesis H0: mu(consition4) = mu(condition3) and alternative hypothesis Ha: mu(consition4) < mu(condition3).

Two-Sample t test for Bedrooms

Assumptions Checking

In order to figure out if the price for a four-bedroom property was significantly higher than a three-bedroom property, the two samples I need to use for the two-sample t test are the subsets of data that has three bedrooms and data that has four bedrooms. The data are continuous in nature (price), and we assume both samples are simple random samples from

their respective populations. Besides, to use the two-sample t-test, it requires either two independent normal populations or two large enough independent samples such that the Central Limit Theorem holds.

```
data1$bedrooms <- factor(data1$bedrooms)
room3 <- data1$price[data1$bedrooms==3]
room4 <- data1$price[data1$bedrooms==4]
l1<-length(room3)
l2<-length(room4)
```

There are 2030 three-bedroom properties and 1531 four-bedroom properties, which indicates the two samples are large enough such that the Central Limit Theorem holds. Additionally, the two samples are independent as properties with three bedrooms should not have relative with the properties with four bedrooms. Choosing a three-bedroom properties does not affect the chance of choosing another four-bedroom property. As a result, I am able to use two-sample t test for this question.

Two-Sample t test

```
p1<-t.test(room4,room3, mu=0, alternative="greater")$p.value
```

The p-value, $3.0658932 \times 10^{-47}$, of this two-sample t test is much lower than the alpha I pick, 0.05. So I have enough evidence to conclude that the true mean prices are not the same, and true mean price properties with four-bedrooms is signficant higher than the true mean price of three-bedrooms properties.

Two-Sample t test for Conditions

Check Assumptions

The two samples are properties with condition 4 and properties with condition 3. As I have two samples and do not know the population standard deviations, I use two sample t test. The data are continuous in nature (price), and I assume both samples are simple random samples from their respective populations. Besides, to use the two-sample t-test, it requires either two independent normal populations or two large enough independent samples such that the Central Limit Theorem holds. As I do not know the distributions of the samples, I need to show these two samples are large enough.

```
data1$condition <- factor(data1$condition)
cond3 <- data1$price[data1$condition==3]
cond4 <- data1$price[data1$condition==4]
l3<-length(cond3)
l4<-length(cond4)
```

There are 2874 properties with condition 3 and 1251 properties with condition 4, which indicates the samples sizes are large enough, so the Central Limit Theorem holds. Additionally, the two samples are independent as properties with condition 3 should not have relative with the properties with condition 4. Choosing a properties with condition 3 does not affect the chance of choosing another property with condition 4. As a result, I am able to use two-sample t test for this question.

Two-sample t test

```
p2<- t.test(cond4,cond3, mu=0, alternative="less")$p.value
```

The p-value, 0.0028635, is less than 0.05 (alpha), so I reject the null hypothesis. There is sufficient evidence to show that the true mean are not the same, and the true mean price of properties with condition 3 is more expensive than that of properties with condition 4.

Tests Outcomes

Based on the rejection of both two-sample t tests, I can conclude that true mean prices for properties with three bedrooms and four bedrooms were not the same, and true mean price properties with four-bedrooms was signficant higher than the true mean price of three-bedrooms properties. Also, the true mean were not the same for properties with condition 3 and condition 4, and the true mean price of properties with condition 3 was more expensive than that of properties with condition 4.

Conclusion

For properties sold in Washington State in 2014, the properties with 4 bedrooms were normally much nore expensive than the those with 3 bedrooms. The properties with condition 4, however, were not normally more expensive than the those with condition 3 properties. In contrast, properties with condition 3 are normally much more expensive that those with condition 4.

References

1. https://www.kaggle.com/shree1992/housedata
2. Other are all from class notes.