Statistical Methods Final Project

Ben Marlow, Yilin Chen, Jingjing Yan

12/16/2021

1 Austin, TX house listings

For the beginning stages of this project, we'll be examining the housing market in Austin, Texas contained in the austinHousingData.csv file on the Kaggle website. The data consists of 15171 observations of 27 variables. The variable `city` is a character that can be converted to a factor with 8 levels representing the name of a city or town in or surrounding Austin, Texas. The variable `yearBuilt` is a discrete integer variable that records the year the house was built dating all the way back to 1905. The variable `zipcode` is an integer that can also be converted to a factor with 48 levels that represent the postal code of the property. The variable `homeType` is converted to a factor with 10 levels grouping the housing market into home types. The variable `numOfAppliances` records the number of appliances for each house listing in the data set. Initial exploration shows that the largest variability in the number of appliances occurs across Multiple Occupancy home types. Other variability in the number of appliances occurs when faceting by city.

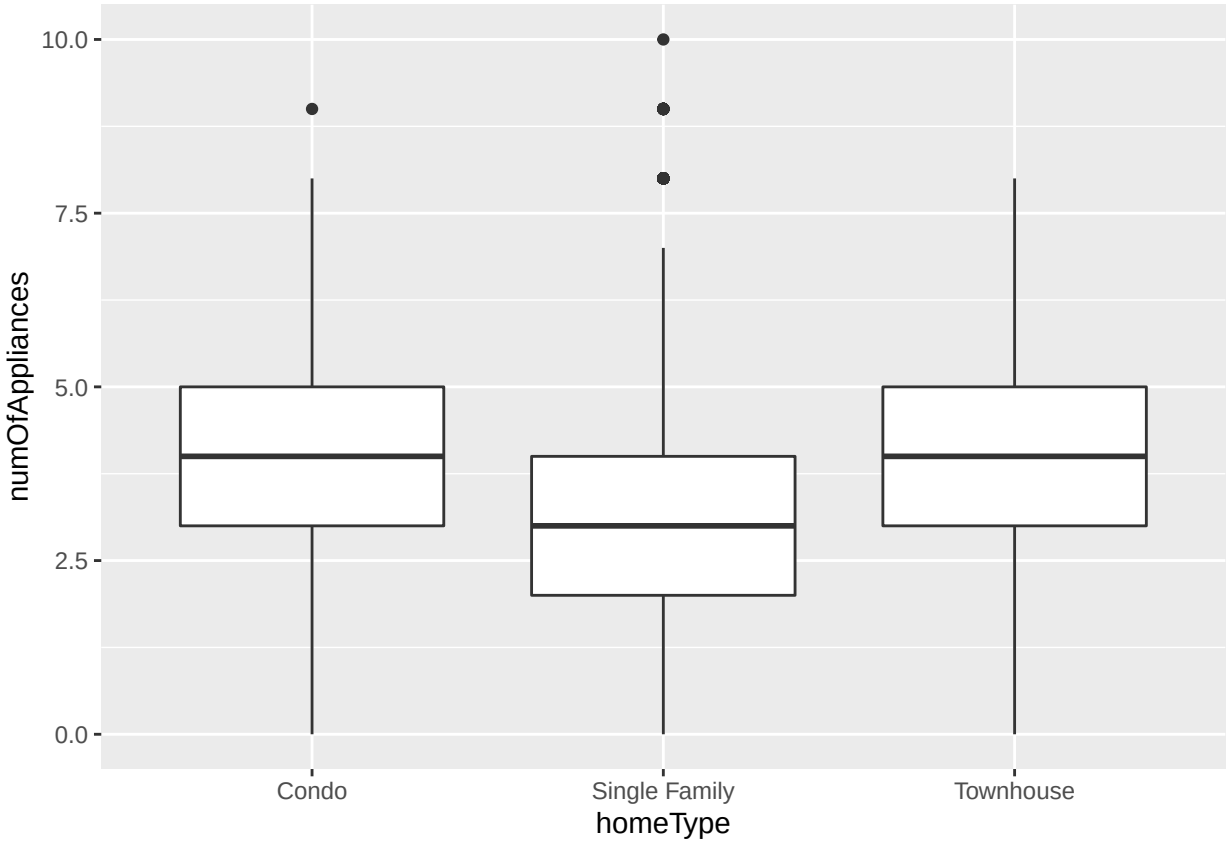
1.1 Question 1

First, is there a significant difference in the number of appliances by home type in Austin? If so, which types vary the most? Before running the test, we'll only consider the top three home types based on count. To test this, we'll perform an ANOVA to see if at least one home type has a significantly different average number of appliances. If there's at least one home type that is significantly different, we can follow up with a multiple comparison test. The ANOVA test resulted in a test statistic F = 7.268 on 2 and 4.949×10^3 degrees of freedom. The p-value for the test is 7.05e-4 which is less than the significance level of 0.05 so we reject the null hypothesis and conclude that there is sufficient evidence to claim that at least one home type in the filtered set experiences a different mean number of appliances.

- State the hypotheses
  - H_0: $\mu_1 = \mu_2 = \mu_3$
  - H_A: at least one home type's average number of appliances is different

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## homeType       2     53  26.262   7.268 0.000705 ***
## Residuals   4949  17883   3.614
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Visualize the data



Because we found that at least one home type was significantly different in the mean number of appliances, we'll now perform a multiple comparison test. Based on the results shown in the table below, there is evidence that Condos have a significantly higher mean number of appliances than Single Family homes because the multiple comparison adjusted p-value for the two is less than 0.05. There is not sufficient evidence to claim that any of the other home types show a significant difference in number of appliances.

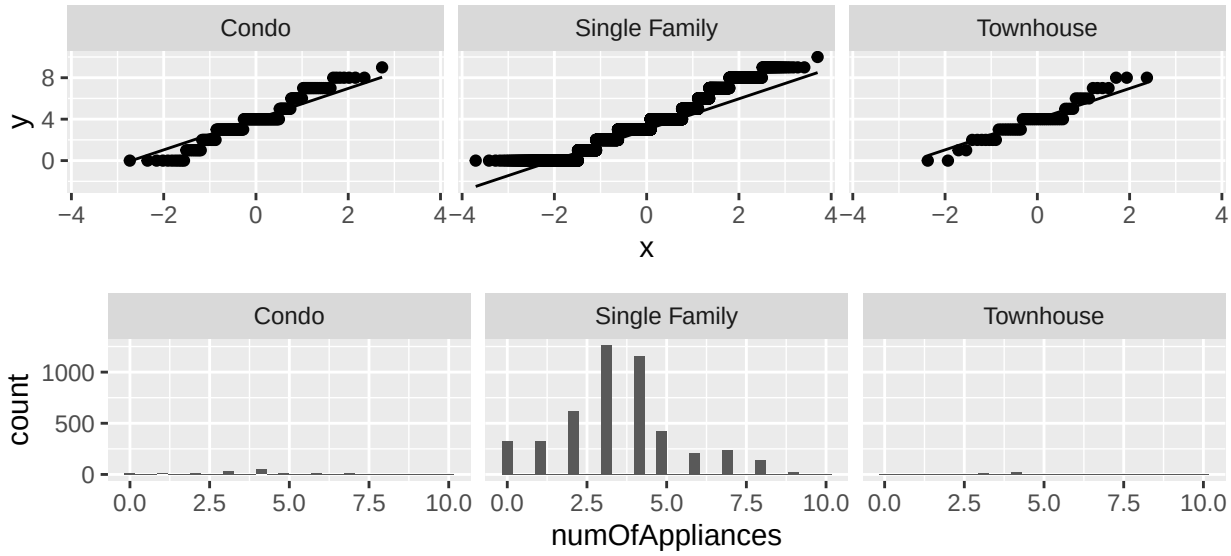
- Which home type has a different mean number of appliances?

```
## # A tibble: 3 x 2
##   homeType      mean_numOfAppliances
##   <fct>                        <dbl>
## 1 Condo                         3.98
## 2 Single Family                 3.48
## 3 Townhouse                     4

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  new_austin_housing1$numOfAppliances and new_austin_housing1$homeType
##
##               Condo  Single Family
## Single Family 0.0035 -
## Townhouse     1.0000 0.1188
##
## P value adjustment method: bonferroni
```

- Checking assumptions – QQ plot and histograms

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## [1] 2
```

The assumptions appear to be met (remember there's a massive discrepancy in sample sizes between home types even after sub-setting).

1.2 Question 2

Second, we want to discover which variables may be driving the housing price. To test this, we'll perform a multiple regression on the `latestPrice` variable. We're going to model the `austin_housing` latest price using the year in which the house was built (`yearBuilt`), living area in square footage (`livingAreaSqFt`), number of price changes (`numPriceChanges`), and number of appliances (`numOfAppliances`).

- State the hypotheses
  - H_0: none of the predictor variables are significant in determining `latestPrice`
  - H_A: one or more of the predictor variables are significant in determining `latestPrice`

```
##
## Call:
```

```
## lm(formula = latestPrice ~ yearBuilt + livingAreaSqFt + numPriceChanges +
##     numOfAppliances, data = new_austin_housing)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -17335323   -142626    -71956     33847  12161446
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.577e+06  3.119e+05  11.470  < 2e-16 ***
## yearBuilt       -1.728e+03  1.577e+02 -10.957  < 2e-16 ***
## livingAreaSqFt   1.612e+02  2.465e+00  65.409  < 2e-16 ***
## numPriceChanges -1.861e+03  1.314e+03  -1.416 0.156713
## numOfAppliances  5.992e+03  1.723e+03   3.477 0.000509 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 399100 on 15165 degrees of freedom
## Multiple R-squared:  0.2246, Adjusted R-squared:  0.2244
## F-statistic:  1098 on 4 and 15165 DF,  p-value: < 2.2e-16
```

What stands out?

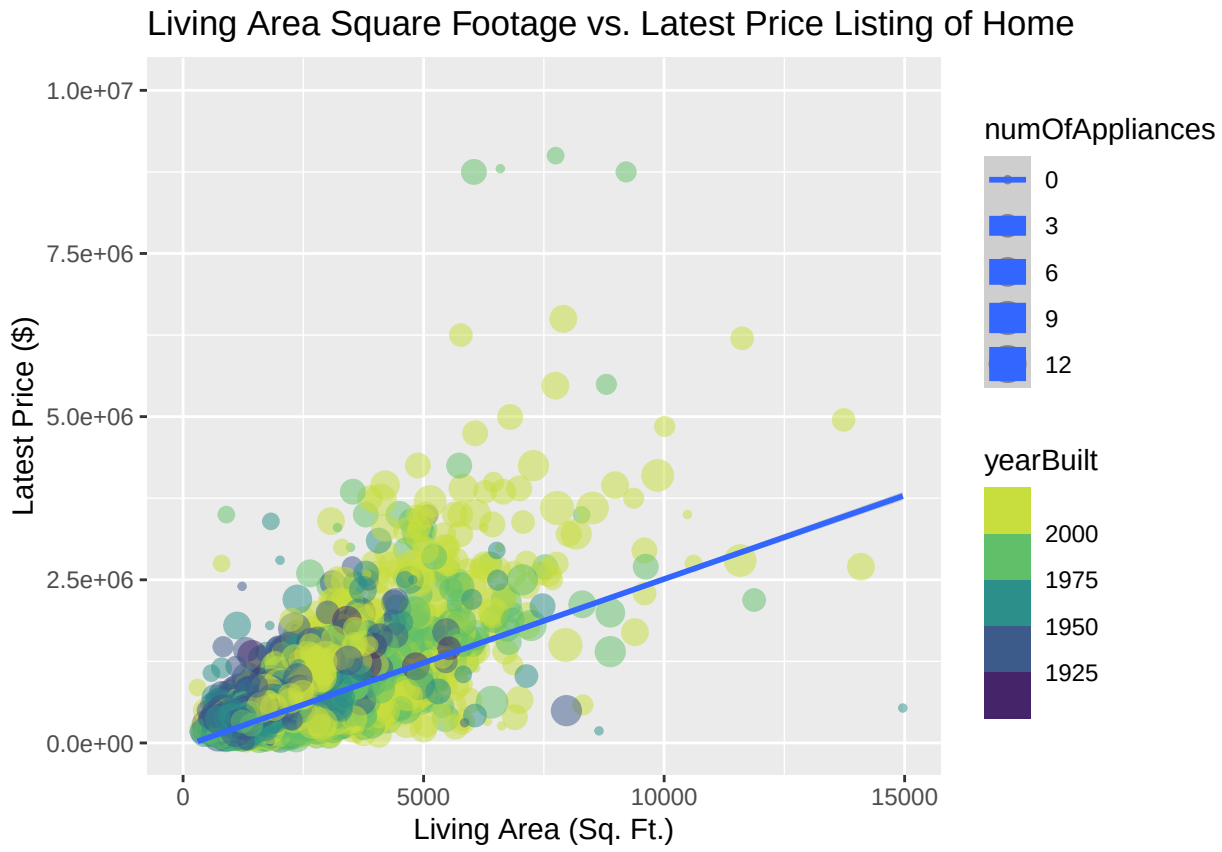
- Overall, the regression is significant at the 0.05 alpha level.
  - F statistic for the overall regression is $F = 1098$ with a p-value < 2.2e-16
- The significant slope variables in this model are `yearBuilt`, `livingAreaSqFt`, and `numOfAppliances`.

t_{test} for the slope of yearBuilt is -10.957 with a p-value < 2.2e-16.

t_{test} for the slope of livingAreaSqFt is 65.409 with a p-value < 2.2e-16.

t_{test} for the slope of numOfAppliances is 3.477 with a p-value of $p = 0.000509$.

```
## `geom_smooth()` using formula 'y ~ x'
```



- What did we learn? Based on this model (using variables `livingAreaSqFt`, `yearBuilt`, and `numOfAppliances`)
  - The latest price will, on average, increase by $161 for each additional square foot of living space.
  - The latest price will, on average, decrease by $1,728 as the year in which the house was built increases by 1.
  - The latest price will, on average, increase by $5,992 for each additional appliance in the home.

Now we're going to find the best model using different combinations of the statistically significant variables.

Compare the full model (fit) to models with a combination of only 2 predictor variables.

```
# Currently the best model (3 predictors)
fit <- lm(latestPrice ~ yearBuilt + livingAreaSqFt + numOfAppliances, data = new_austin_housing)

# First 2 variable model with yearBuilt and livingAreaSqFt
fit_a <- lm(latestPrice ~ yearBuilt + livingAreaSqFt, data = new_austin_housing)

# Second 2 variable model with numOfAppliances and livingAreaSqFt
fit_b <- lm(latestPrice ~  numOfAppliances + livingAreaSqFt, data = new_austin_housing)

# Third 2 variable model with yearBuilt and numOfAppliances
fit_c <- lm(latestPrice ~ yearBuilt  + numOfAppliances, data = new_austin_housing)
```

- 3 predictor model (fit): $R^2_{adj} = 0.2244$

- First 2 predictor model (fit_a): $R^2_{adj} = 0.2238$
- Second 2 predictor model (fit_b): $R^2_{adj} = 0.2184$
- Third 2 predictor model (fit_c): $R^2_{adj} = 0.004196$

R^2_{adj} is highest for the 3 predictor model. Thus, "fit" will be our preferred regression.

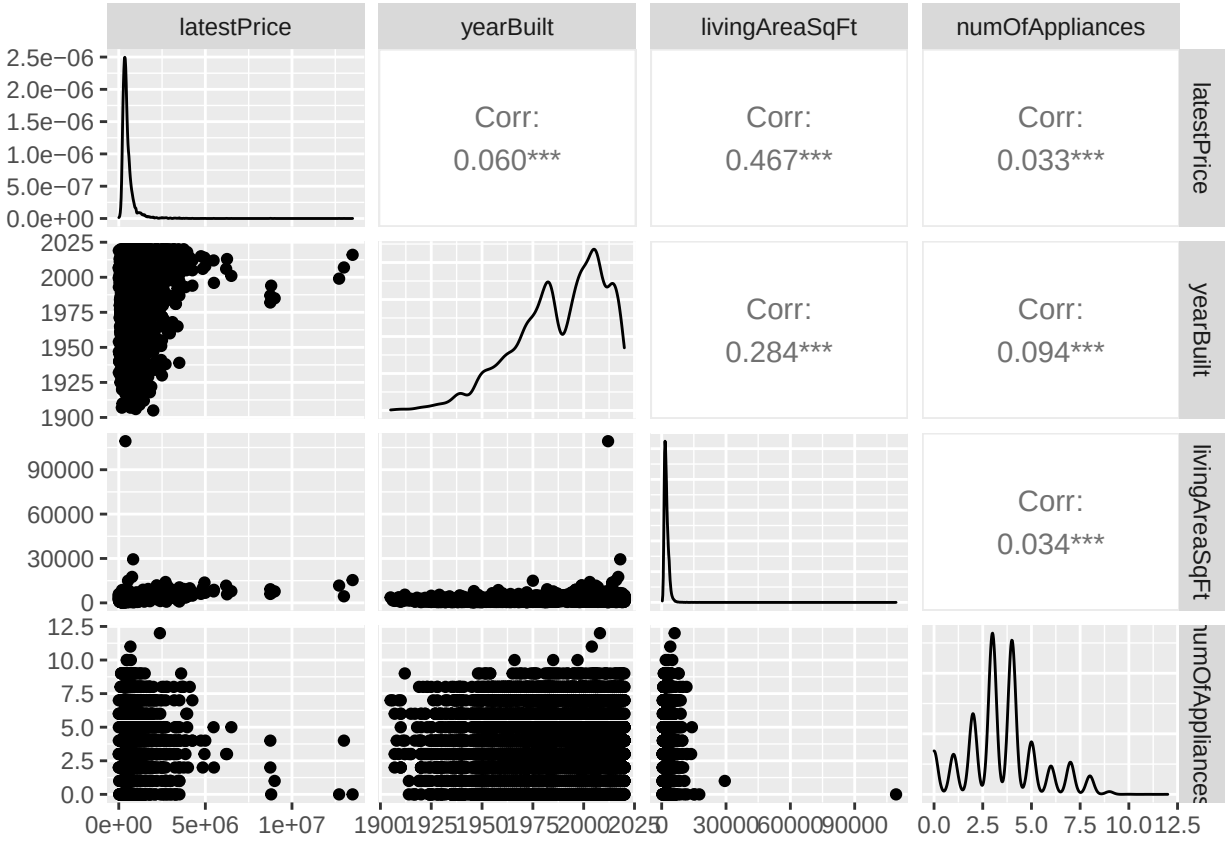
Sample estimate regression equation:

$\hat{y}$ = 3.539e+06 + (-1.710e+03) * x_1 + 1.609e+02 * x_2 + 5.827e+03 * x_3

x_1: yearBuilt

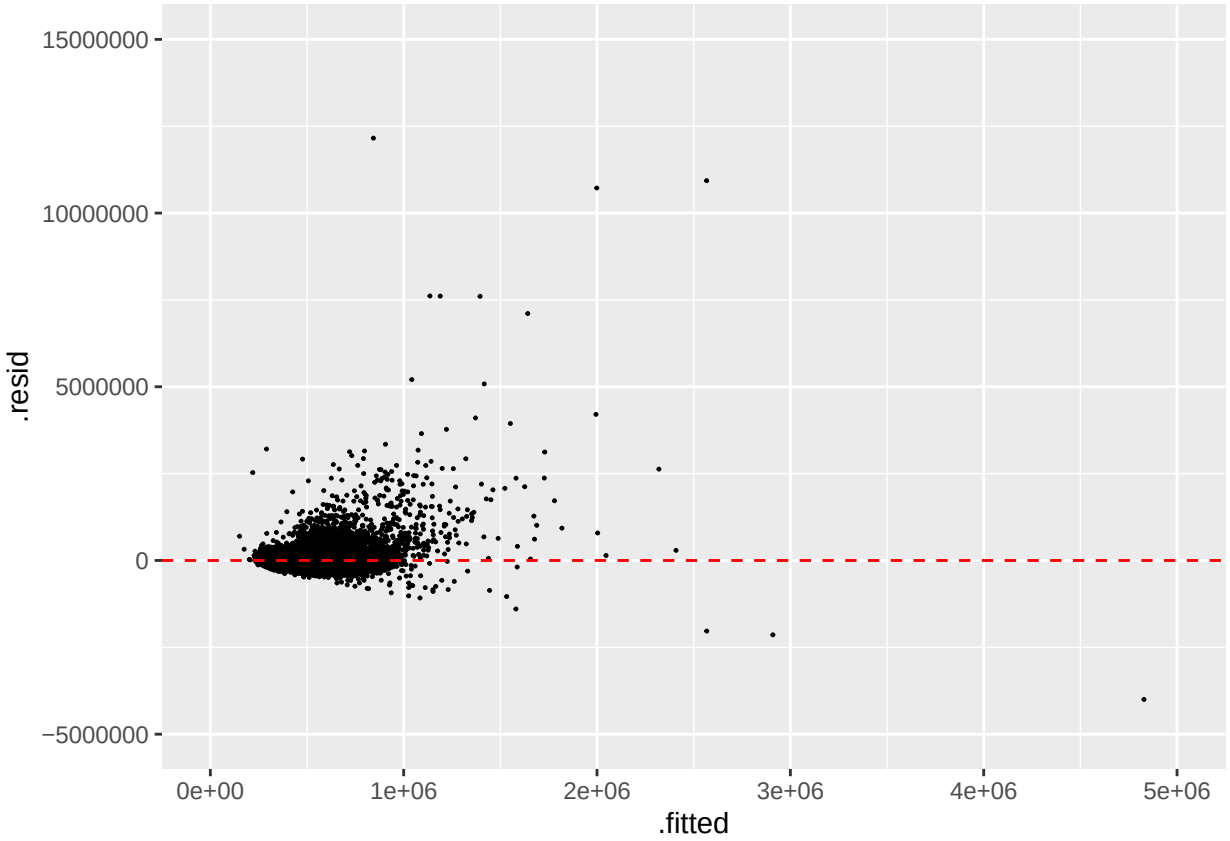
x_2: livingAreaSqFt

x_3: numOfAppliances



Residual model checking

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



- The residual plot exhibits a somewhat noticeable "fanning" pattern towards the top. Hence, we can assume the distribution of latest price does not follow a normal shape.

1.3 Question 3

The third question we'll be answering for this data set is whether home type and zip code are independent. We'll perform a chi-squared test of independence using a two way-table. We'll set $\alpha = 0.05$. To narrow down our study, we'll only keep zip codes with 500 or more observations.

- 1) State the hypothesis
  - H_0: home type and zip code are independent
  - H_a: home type and zip code are dependent

Calculate Expected Counts

```
##                          78704        78717        78723        78732
## Apartment             1.57114806   1.14831784   1.24178557   1.32857704
## Condo                16.82052625  12.29375573  13.29441026  14.22358948
```

```
## Mobile / Manufactured   0.27726142   0.20264433   0.21913863   0.23445477
## MultiFamily             0.09242047   0.06754811   0.07304621   0.07815159
## Multiple Occupancy      3.97408038   2.90456866   3.14098704   3.36051839
## Other                   0.36968190   0.27019243   0.29218484   0.31260636
## Residential             1.75598900   1.28341406   1.38787799   1.48488022
## Single Family         669.40149234 489.25094908 529.07370075 566.05197015
## Townhouse               9.24204739   6.75481084   7.30462102   7.81515905
## Vacant Land             2.49535279   1.82379893   1.97224768   2.11009294
##                              78737        78739        78745        78748
## Apartment               1.37308548   1.34638042   2.3122136    2.5436575
## Condo                  14.70009164  14.41419034  24.7542872   27.2320984
## Mobile / Manufactured   0.24230920   0.23759654   0.4080377    0.4488807
## MultiFamily             0.08076973   0.07919885   0.1360126    0.1496269
## Multiple Occupancy      3.47309857   3.40555046   5.8485404    6.4339573
## Other                   0.32307894   0.31679539   0.5440503    0.5985077
## Residential             1.53462495   1.50477811   2.5842388    2.8429114
## Single Family         585.01518523 573.63725619 985.1390234 1083.7477419
## Townhouse               8.07697343   7.91988480  13.6012567   14.9626915
## Vacant Land             2.18078282   2.13836890   3.6723393    4.0399267
##                             78749        78757        78759
## Apartment               1.7224768   1.16389580   1.24846184
## Condo                  18.4406336  12.46053148  13.36588559
## Mobile / Manufactured   0.3039665   0.20539338   0.22031680
## MultiFamily             0.1013222   0.06846446   0.07343893
## Multiple Occupancy      4.3568530   2.94397172   3.15787407
## Other                   0.4052887   0.27385783   0.29375573
## Residential             1.9251211   1.30082472   1.39533970
## Single Family         733.8764236 495.88807436 531.91818301
## Townhouse              10.1322163   6.84644587   7.34389318
## Vacant Land             2.7356984   1.84854038   1.98285116
```

- 2) Compute the test statistic and p-value

```
chisq.test(O)
```

```
## Warning in chisq.test(O): Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  O
## X-squared = 781.84, df = 90, p-value < 2.2e-16
```

- 3) State the decision
  - p-value $< \alpha$: Reject H_0.
  - 2.2e-16 < 0.05: Reject H_0.
- 4) Interpret the decision
  - Because the p-value is less than the significance level α, we reject the null hypothesis and conclude there is sufficient evidence to claim that home type and zip code are dependent.
- 5) Check assumptions

a) Data in the table cells are frequencies or counts - Yes

b) The levels (or categories) of the variables are mutually exclusive - Yes

c) There are 2 variables, and both are measured as categories - Yes

2 World Climate data analysis

Using data available on Kaggle, we're going to investigate climate across three regions around the world (Western Europe - Ireland, Central Europe - Italy, and Eastern Asia - Tokyo, Japan). The variables in the weather data include the date, time, temperature, wind speed, humidity, pressure, and a brief description. Weather records were extracted from the open source Rest API. There are 15 variables in each data set - we'll use roughly half of them when conducting our analysis.

2.1 Question 1:

What is the 95% confidence interval for proportion of `overcast clouds` weather in each country/region?

```
##
##  1-sample proportions test with continuity correction
##
## data:  ireland_n_coulds$n_clouds[11] out of sum(ireland_n_coulds$n_clouds), null probability 0.5
## X-squared = 3318.4, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.1780674 0.1948042
## sample estimates:
##         p
## 0.1862919
```

To test this question, we first perform the single proportion of `overcast clouds` for Ireland. The number of predictions of `overcast clouds` weather is 1571, while the total number of weather `Description` is 8433. The proportion test with a confidence level of 95% results in a rejection of the null hypothesis of the true p equal to 0.5 giving a p-value $< 2.2e-16$. According to the proportion test, we are 95% confident that the true proportion of `overcast clouds` weather `Description` that occurs in the region of Ireland is between 17.8% and 19.5%.

```
##
##  1-sample proportions test with continuity correction
##
## data:  italy_n_coulds$n_clouds[12] out of sum(italy_n_coulds$n_clouds), null probability 0.5
## X-squared = 1860.2, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2349973 0.2550688
## sample estimates:
##         p
## 0.2448951
```

Similarly, we now perform the single proportion of `overcast clouds` for Italy. The number of predictions of `overcast clouds` weather is 1751, while the total number of weather `Description` is 7150. The proportion test with a confidence level of 95% results in a rejection of the null hypothesis of the true p equal to 0.5 giving a p-value $< 2.2e-16$. According to the proportion test, we are 95% confident that the true proportion of `overcast clouds` weather `Description` that occurs in the region of Italy is between 23.5% and 25.5%.

```
##
##  1-sample proportions test with continuity correction
```

```
##
## data:  tokyo_n_coulds$n_clouds[12] out of sum(tokyo_n_coulds$n_clouds), null probability 0.5
## X-squared = 6371.7, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.02804010 0.03626896
## sample estimates:
##          p
## 0.03190319
```

Third, we perform the single proportion of `overcast clouds` for Tokyo. The number of predictions of `overcast clouds` weather is 232, while the total number of weather `Description` is 7272. The proportion test with a confidence level of 95% results in a rejection of the null hypothesis of the true p equal to 0.5 giving a p-value $< 2.2e-16$. According to the proportion test, we are 95% confident that the true proportion of `overcast clouds` weather `Description` that occurs in the region of Tokyo is between 2.8% and 3.6%.

Therefore, we can conclude that comparing with these two European regions, Tokyo of Japan experiences the smallest proportion of overcast clouds weather.

Does the proportion of overcast weather vary across two of the European countries (i.e., Ireland and Italy) in this certain period?

In the real world, the proportion of `overcast clouds` weather can have an impact on people's daily lives such as solar panel generation, precipitation prediction, and even the human mood. Thus, it is necessary to clarify the regional representative ability of the `overcast clouds` weather at a given location.

Therefore, taking the example of the two European countries, we tried to clarify if the proportion of overcast weather are varied. We perform the two sample proportion test. The two sample proportion test with a confidence level of 95% results in a rejection of the null hypothesis giving a p-value = 0.0003648 < 0.05. According to the proportion test, we are 95% confident that the proportion of `overcast clouds` weather `Description` that occurs in Ireland is significantly different than that of Italy.

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(ireland_n_coulds$n_clouds[11], italy_n_coulds$n_clouds[12]) out of c(sum(italy_n_coulds$n_cl
## X-squared = 12.705, df = 1, p-value = 0.0003648
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.03901184 -0.01133781
## sample estimates:
##    prop 1    prop 2
## 0.2197203 0.2448951
```

2.2 Question 2:

What are the primary linear factors, including Temperature, Wind_speed, pressure, humidity, and visibility, that impact the "feels like" temperature, taking the example of Tokyo Japan?

```
##
## Call:
## lm(formula = feels_like ~ Temperature + Wind_speed + pressure +
##     humidity + visibility, data = tokyo_weather)
##
```

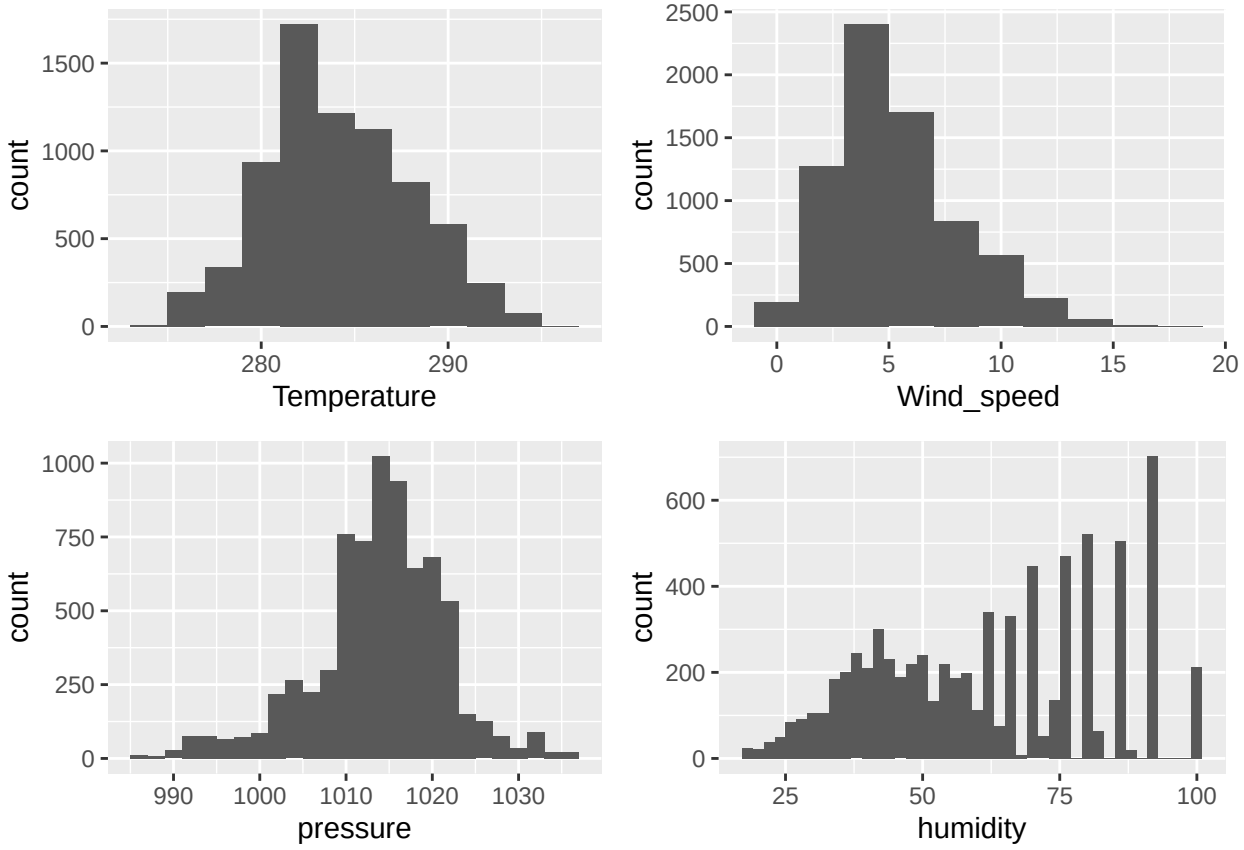
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73962 -0.12301 -0.01409  0.08327  1.10228
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -5.267e+01  4.826e-01 -109.141  < 2e-16 ***
## Temperature  1.175e+00  7.265e-04 1617.129  < 2e-16 ***
## Wind_speed  -6.902e-01  9.560e-04 -721.999  < 2e-16 ***
## pressure    -1.016e-03  3.669e-04   -2.769  0.00563 **
## humidity     4.329e-02  1.413e-04  306.402  < 2e-16 ***
## visibility  -2.092e-06  1.479e-06   -1.414  0.15727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2182 on 7266 degrees of freedom
## Multiple R-squared:  0.998,  Adjusted R-squared:  0.998
## F-statistic: 7.252e+05 on 5 and 7266 DF,  p-value: < 2.2e-16
```

To address this question, we apply multiple linear regression on the `feels like` temperature in Tokyo. The regression of the `feels like` against the multiple variables of `Temperature`, `Wind_speed`, `pressure`, `humidity`, and `visibility` results in a rejection of the null hypothesis of no linear correlation with a test statistic $\mathrm{F} = 7.252e+05$ on 5 and 7266 degrees of freedom giving a p-value $< 2.2e-16$. Therefore, we reject the null hypothesis of no linear relationships and conclude there is a statistically significant linear relationship in the dependent variable of `feels like` with the multiple independent variables (i.e., `Temperature`, `Wind_speed`, `pressure`, `humidity`, and `visibility`). The estimate of the `Temperature` slope is 1.175 which suggest that, on average, `feels like` temperature increases 1.175 K per K increase of the air `Temperature`. The estimate of the `Wind_speed` slope is -0.6902 which suggest that, on average, `feels like` temperature decreases 0.6902 K per m/s increase of the `Wind_speed`. The estimate of the `pressure` slope is -0.001016 which suggest that, on average, `feels like` temperature decreases 0.001016 K per hPa increase of the `pressure`. The estimate of the `humidity` slope is 0.04329 which suggest that, on average, `feels like` temperature increases 0.04329 K per unit increase of the `humidity`. The estimate of the `visibility` slope is -0.000002092 with the p_value $= 0.15727 > 0.05$, suggesting that `visibility` is not a significant factor that influences the `feels like` temperature. Based on the $R^2 = 0.998$ value, the linear regression model using these full 5 predictors explains 99.8% of the variation in feels like.

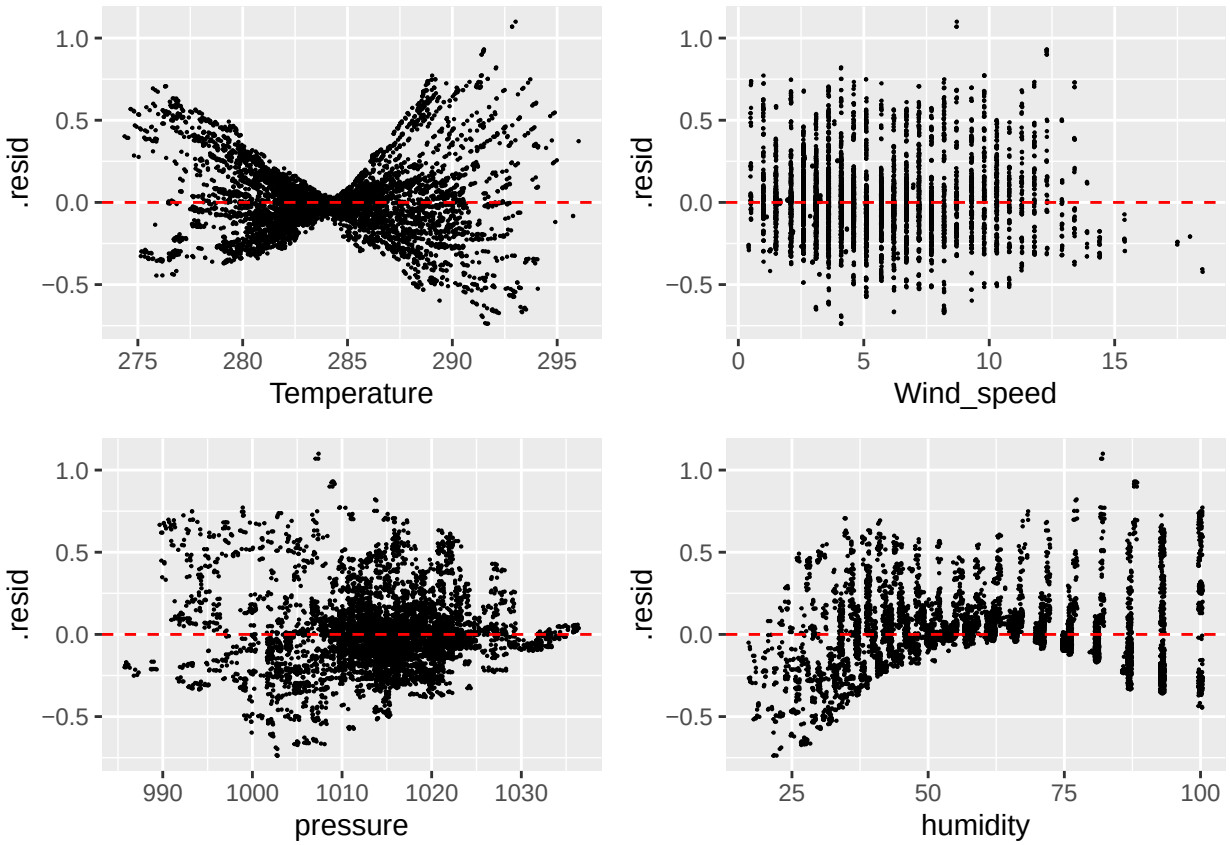
```
##
## Call:
## lm(formula = feels_like ~ Temperature + Wind_speed + pressure +
##     humidity, data = tokyo_weather)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73860 -0.12268 -0.01489  0.08278  1.09950
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -5.263e+01  4.816e-01 -109.273  < 2e-16 ***
## Temperature  1.175e+00  7.265e-04 1617.296  < 2e-16 ***
## Wind_speed  -6.902e-01  9.560e-04 -721.949  < 2e-16 ***
## pressure    -1.089e-03  3.633e-04   -2.998  0.00273 **
## humidity     4.336e-02  1.330e-04  326.072  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2183 on 7267 degrees of freedom
## Multiple R-squared:  0.998,  Adjusted R-squared:  0.998
## F-statistic: 9.063e+05 on 4 and 7267 DF,  p-value: < 2.2e-16
```

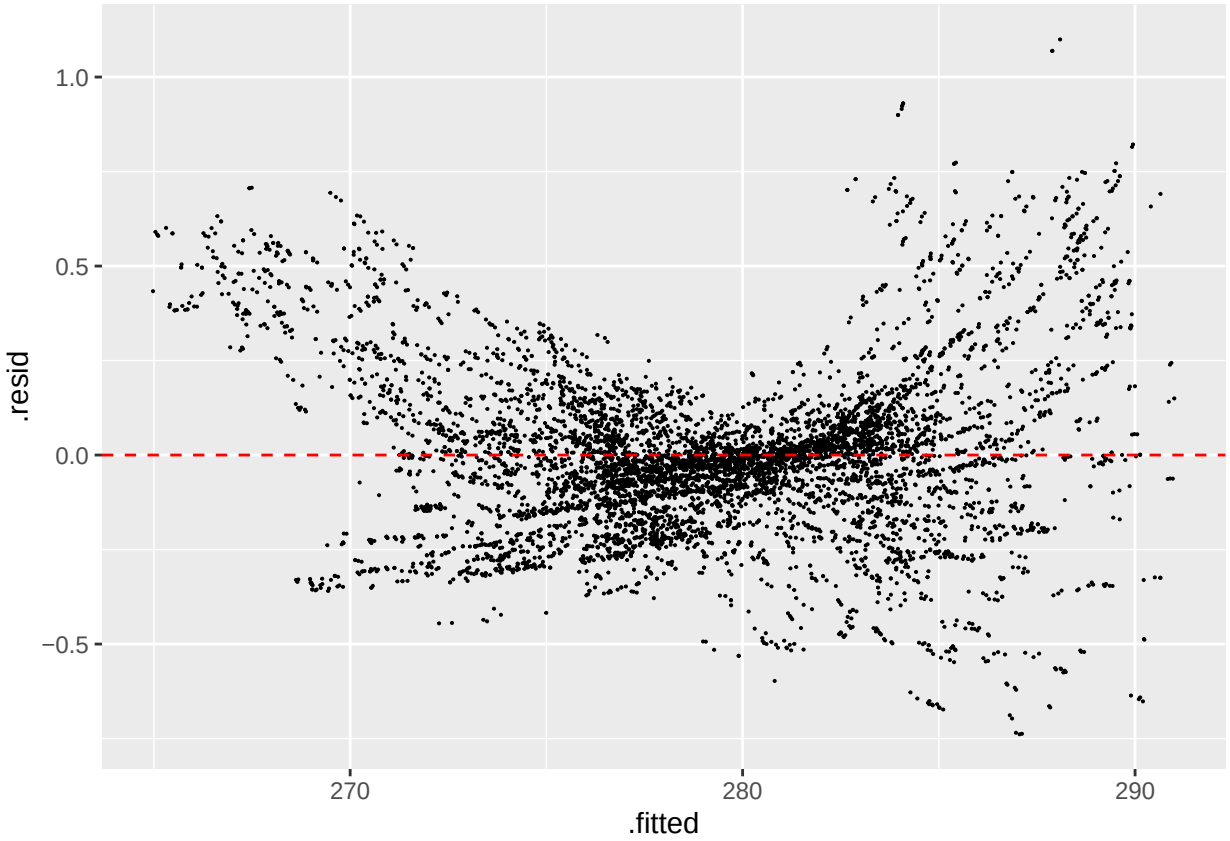
In fact, we can refine the regression model by reducing the insignificant variable `visibility` to obtain the reduced model. The regression of the `feels like` against the multiple variables of `Temperature`, `Wind_speed`, `pressure`, and `humidity` results in a rejection of the null hypothesis of no linear correlation with a test statistic F = 9.063e+05 on 4 and 7267 degrees of freedom giving a p-value $< 2.2e-16$. Based on the $R^2 = 0.998$ value, the linear regression model using these reduced 4 predictors explains 99.8% of the variation in `feels like`. Even though the value of R^2 for both the reduced model and full model is the same which equates to 0.998, we conclude that the reduced model, because it is more simplified, to be the best regression model.



Assumptions check: (1) In this study, we have no reason to believe that the predictors of `Temperature`, `Wind_speed`, `pressure`, `humidity`, and `visibility` won't be independent of each other. (2) According to the histograms above, we can see that nearly all four predictors follow a normal distribution (outside of maybe humidity). (3) Based on the residual plots below, we can find that the residuals across the predictors are about in a constant belt, apart from the slightly binomial distribution of `Temperature`. Therefore, we believe that the assumptions for our multiple linear regression model are met.

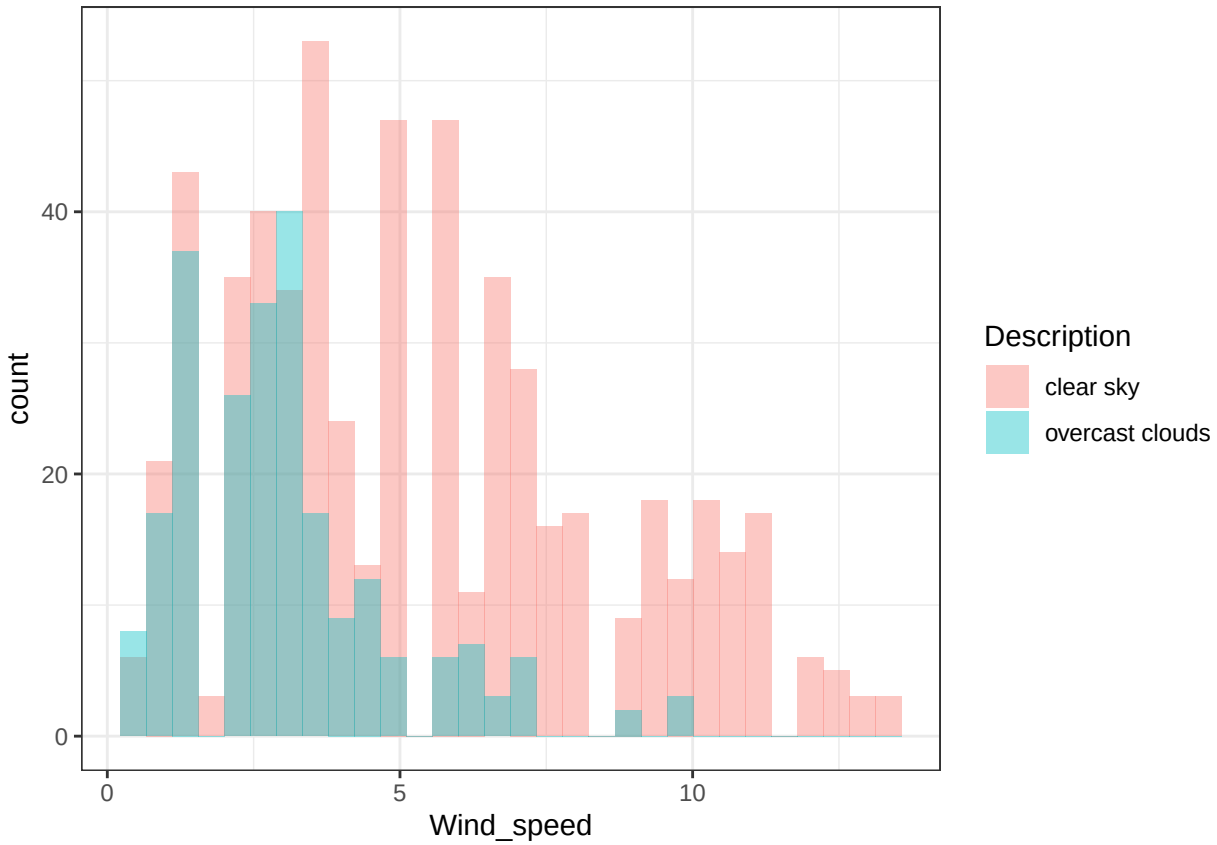

.resid
1.0
0.5
0.0
−0.5
275
280
285
290
295
Temperature
0
5
10
15
Wind_speed
990
1000
1010
1020
1030
pressure
25
50
75
100
humidity

The residuals vs. fitted diagnostics is plotted below. The `feels_like` temperature is more stable and concentrated around 280 K, while below or above this value, the `feels_like` temperature experiences the larger range, which means the weather conditions are harder to predict.



2.3 Question 3:

Whether the average of `Wind_speed` in `clear sky` weather is greater than that in `overcast clouds` weather in Tokyo?




```
##
##  Welch Two Sample t-test
##
## data:  clearsky_wind$Wind_speed and overcast_wind$Wind_speed
## t = 13.028, df = 702.98, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.009983      Inf
## sample estimates:
## mean of x mean of y
##  5.367318  3.066466
```

We can perform a two-sample t-test to test for a difference in the average of Wind_speed in `overcast clouds` and `clear sky` weather in Tokyo. The test for the average `Wind_speed` in different weather conditions (`overcast clouds` and `clear sky`) rejects the null hypothesis that the average `Wind_speed` in `clear sky` weather is the same or smaller than that in the `overcast clouds` weather with a test statistic t = 13.028 with 702.98 degree of freedom giving a p-value $< 2.2e-16$. Therefore we reject the null hypothesis of no difference in the average `Wind_speed` under weather conditions of `overcast clouds` and `clear sky`. The sample estimated mean of `Wind_speed` in `clear sky` weather is 5.367318 but in `overcast clouds` weather equals to 3.066466. We cannot conclude definitively that the atmosphere is more stable in `overcast clouds` weather condition from the data, however, there is strong evidence that the average `Wind_speed` is greater in `clear sky` weather than `overcast clouds`.

3 Conclusion

Dataset 1:

We found that at least one home type's average number of appliances is different. Condos have a significantly higher mean number of appliances than Single Family homes.

Square footage of living space and total number of appliances drive the housing price. The later the year the house is built the more likely the listing price is to decrease (on average). The number of price changes has little to no effect on housing price.

We found home type and zip code are dependent.

Dataset 2:

European countries, Ireland and Italy, experience more `overcast clouds` weather than the Asian region of Tokyo, but when observing the two European countries specifically, they don't have a similar proportion of `overcast clouds` weather. This means the number of observations of `overcast clouds` weather in one place (Italy or Ireland) doesn't necessarily mirror the rest of the region or continent for that matter (Europe).

The `feels like` temperature can be significantly impacted by `Temperature`, `Wind_speed`, `pressure`, and `humidity`. Around the value of 280 K, the `feels_like` temperature is more stable, but below or beyond that, the residuals generate more variability which means weather conditions are tougher to gauge.

Based on our study of Tokyo, Japan, we found that that the average `Wind_speed` is greater in `clear sky` weather than `overcast clouds` weather.

4 Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(here)
library(ggplot2)
library(tidyverse)
library(readxl)
library(dplyr)
library(GGally)
library(broom)
source(here("Final draft/R", "multiplot.R"))
austin_housing <- read.csv(here::here("data", "austinHousingData.csv"))
ireland_weather <- read.csv(here::here("data", "ireland_weather.csv"))
italy_weather <- read.csv(here::here("data", "italy_weather.csv"))
tokyo_weather <- read.csv(here::here("data", "tokyo_weather.csv"))
house_vars <- c("city", "zipcode", "latitude", "longitude", "propertyTaxRate", "garageSpaces", "homeTyp
new_austin_housing <- austin_housing[house_vars]
new_austin_housing <- new_austin_housing %>% filter(city != 'road')
glimpse(new_austin_housing)
new_austin_housing$homeType <- factor(new_austin_housing$homeType)
levels(new_austin_housing$homeType)
new_austin_housing %>% group_by(homeType) %>% summarize(homeType_totals = n())
new_austin_housing1 <- filter(new_austin_housing, homeType == c("Single Family", "Condo", "Townhouse"))
## perform the ANOVA
model <- aov(numOfAppliances ~ homeType, data = new_austin_housing1)
summary(model)
ggplot(data = new_austin_housing1, mapping = aes(x = homeType, y = numOfAppliances)) + geom_boxplot()
## examine mean salary by position
new_austin_housing1 %>% group_by(homeType) %>% summarize(mean_numOfAppliances = mean(numOfAppliances))

## make pairwise comparisons using Bonferroni correction
pairwise.t.test(new_austin_housing1$numOfAppliances, new_austin_housing1$homeType, p.adjust.method = "b
house_qq <- new_austin_housing1 %>% ggplot(aes(sample = numOfAppliances)) + facet_wrap(~ homeType) + st

house_hist <- new_austin_housing1 %>% ggplot(aes(x = numOfAppliances)) + facet_wrap(~ homeType) + geom_

multiplot(house_qq, house_hist, rows = 2, cols = 1)
fit <- lm(latestPrice ~ yearBuilt + livingAreaSqFt + numPriceChanges + numOfAppliances, data = new_aust
summary(fit)
ggplot(data = new_austin_housing, aes(x = livingAreaSqFt, y = latestPrice, color = yearBuilt, size = nu
geom_point(alpha = .5, position = "jitter") + stat_smooth(method = "lm") + xlim(0, 15000) + ylim(1e+04,
# Currently the best model (3 predictors)
fit <- lm(latestPrice ~ yearBuilt + livingAreaSqFt + numOfAppliances, data = new_austin_housing)

# First 2 variable model with yearBuilt and livingAreaSqFt
fit_a <- lm(latestPrice ~ yearBuilt + livingAreaSqFt, data = new_austin_housing)

# Second 2 variable model with numOfAppliances and livingAreaSqFt
fit_b <- lm(latestPrice ~  numOfAppliances + livingAreaSqFt, data = new_austin_housing)

# Third 2 variable model with yearBuilt and numOfAppliances
fit_c <- lm(latestPrice ~ yearBuilt  + numOfAppliances, data = new_austin_housing)
summary(fit)
```

```
summary(fit_a)
summary(fit_b)
summary(fit_c)
library(GGally)
ggpairs(dplyr::select(new_austin_housing, latestPrice, yearBuilt, livingAreaSqFt, numOfAppliances))
library(broom)
df <- augment(fit)
ggplot(df, aes(x = .fitted, y = .resid)) +
geom_point(size = 0.2) +
geom_hline(yintercept = 0, color = "red", lty = 2) + xlim(0, 5.0e+06) + ylim(-5e+06, 1.5e+07)
new_austin_housing2 <- new_austin_housing %>% group_by(zipcode) %>% filter(n() >= 500)


new_vars <- c("homeType", "zipcode")
new_austin_housing2 <- new_austin_housing2[new_vars]

O <- table(new_austin_housing2)

row_totals <- apply(O, 1, sum)

col_totals <- apply(O, 2, sum)

E <- outer(row_totals, col_totals) / sum(O)
E
chisq.test(O)
ireland_n_coulds <- ireland_weather %>% group_by(Description) %>% summarize(n_clouds = n())
prop.test(x = ireland_n_coulds$n_clouds[11], n = sum(ireland_n_coulds$n_clouds), conf.level = 0.95)
italy_n_coulds <- italy_weather %>% group_by(Description) %>% summarize(n_clouds = n())
prop.test(x = italy_n_coulds$n_clouds[12], n = sum(italy_n_coulds$n_clouds), conf.level = 0.95)
tokyo_n_coulds <- tokyo_weather %>% group_by(Description) %>% summarize(n_clouds = n())
prop.test(x = tokyo_n_coulds$n_clouds[12], n = sum(tokyo_n_coulds$n_clouds), conf.level = 0.95)
prop.test(x = c(ireland_n_coulds$n_clouds[11], italy_n_coulds$n_clouds[12]), n = c(sum(italy_n_coulds$n
fit_full <- lm(feels_like ~ Temperature + Wind_speed + pressure + humidity + visibility,
          data = tokyo_weather)
summary(fit_full)
fit_reduced <- lm(feels_like ~ Temperature + Wind_speed + pressure + humidity, data = tokyo_weather)
summary(fit_reduced)
hist_T <- ggplot(tokyo_weather) +
  geom_histogram(aes(x = Temperature), binwidth=2)
hist_W <- ggplot(tokyo_weather) +
  geom_histogram(aes(x = Wind_speed), binwidth=2)
hist_p <- ggplot(tokyo_weather) +
  geom_histogram(aes(x = pressure), binwidth=2)
hist_h <- ggplot(tokyo_weather) +
  geom_histogram(aes(x = humidity), binwidth=2)
multiplot(hist_T, hist_W, hist_p, hist_h, cols = 2)
resid_T <- ggplot(fit_reduced) +
  geom_point(aes(x = Temperature, y = .resid), size = 0.1, position = "jitter") +
  geom_hline(aes(yintercept = 0), lty = 2, color = "red")
resid_W <- ggplot(fit_reduced) +
  geom_point(aes(x = Wind_speed, y = .resid), size = 0.1, position = "jitter") +
  geom_hline(aes(yintercept = 0), lty = 2, color = "red")
resid_p <- ggplot(fit_reduced) +
```

```
  geom_point(aes(x = pressure, y = .resid), size = 0.1, position = "jitter") +
  geom_hline(aes(yintercept = 0), lty = 2, color = "red")
resid_h <- ggplot(fit_reduced) +
  geom_point(aes(x = humidity, y = .resid), size = 0.1, position = "jitter") +
  geom_hline(aes(yintercept = 0), lty = 2, color = "red")
multiplot(resid_T, resid_W, resid_p, resid_h, cols = 2)
df <- augment(fit_reduced)
ggplot(df, aes(x = .fitted, y = .resid)) +
  geom_point(size = 0.05) +
  geom_hline(yintercept = 0, color = "red", lty = 2)
tokyo_wind_weather <- subset(tokyo_weather, subset = tokyo_weather$Description %in% c("overcast clouds"
ggplot(data = tokyo_wind_weather, aes(x = Wind_speed, fill = Description)) +
  geom_histogram(position = "identity", alpha = 0.4, bins = 30) +
  theme_bw()
overcast_wind <- tokyo_wind_weather %>% filter(Description == "overcast clouds")
clearsky_wind <- tokyo_wind_weather %>% filter(Description == "clear sky")
t.test(clearsky_wind$Wind_speed, overcast_wind$Wind_speed, mu = 0, alternative = "greater")
```