Week8_AdamsKimberly

2022-07-30

Housing Dataset

1) Explain any transformations or modifications you made to the dataset

- When importing the data, I replaced spaces with "." in the column names
- I did not transform the data as often the transformation causes more problems. There is a potential for transforming variables such as sale price to try to normalize the data spread from a positive skew to a more normal distribution, but I did not do this.

2) Create two models; one that will contain the variables Sale Price and Square Foot of Lot and one that will contain Sale Price and several additional predictors of your choice. Explain the basis for your additional predictor selections.

Creating linear models for:

- Price and lot size = how much am I paying for the land space?
- Price and Year Built and lot size = Are older houses cheaper and bigger lots more expensive?

```
PriceBySqFtLot.lm = lm(Sale.Price ~ sq_ft_lot, data = housing_data)

PriceBuildLotSize.lm = lm(Sale.Price ~ sq_ft_lot + year_built, data = housing_data)
```

3) Execute a summary() function on two variables defined in the previous step to compare the model results.

```
summary(PriceBySqFtLot.lm)
```

```
##
## Call:
## lm(formula = Sale.Price ~ sq_ft_lot, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2016064  -194842   -63293    91565  3735109
##
## Coefficients:
##                  Estimate   Std. Error t value       Pr(>|t|)
## (Intercept) 641821.40609   3799.91526  168.90 <0.0000000000000002 ***
## sq_ft_lot        0.85099      0.06217   13.69 <0.0000000000000002 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 401500 on 12863 degrees of freedom
## Multiple R-squared:  0.01435,    Adjusted R-squared:  0.01428
## F-statistic: 187.3 on 1 and 12863 DF,  p-value: < 0.00000000000000022
```

```
summary(PriceBuildLotSize.lm)
```

```
##
## Call:
## lm(formula = Sale.Price ~ sq_ft_lot + year_built, data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2441124  -166193   -48805    74921  3634286
##
## Coefficients:
##                     Estimate    Std. Error t value          Pr(>|t|)
## (Intercept) -11702314.87772  399110.28901  -29.32 <0.0000000000000002 ***
## sq_ft_lot           1.10362       0.06054   18.23 <0.0000000000000002 ***
## year_built       6190.92039     200.15623   30.93 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 387400 on 12862 degrees of freedom
## Multiple R-squared:  0.08259,    Adjusted R-squared:  0.08245
## F-statistic:   579 on 2 and 12862 DF,  p-value: < 0.00000000000000022
```

a) What are the R2 and Adjusted R2 statistics?

Single model with just lot size has a R2 of 0.014 while the model with 2 explanatory variables has an adjusted R2 of 0.083.

b) Explain what these results tell you about the overall model.

The model is not very good at predicting the sale price as it explains less than 10% of the variation within the sale price data.

c) Did the inclusion of the additional predictors help explain any large variations found in Sale Price

The model improved 8x with the addition of the year_built variable, but is still not a great model for this data as it only explains about 8% of the data.

4) Considering the parameters of the multiple regression model you have created.

a) What are the standardized betas for each parameter and what do the values indicate?

```
##  sq_ft_lot year_built
##  0.1553802  0.2636342
```

The standardized beta value is an indication of the how the outcome changes (in units of standard deviations) if the predictor changes by 1 standard deviation. In other words, it gives us an indicator of the importance of the predictor.

The value of year_built is 0.16 and the sq_ft_lot is 0.26 This means that the square footage of the lot is a better predictor than the year the house was built as it has a bigger influence on the Sale Price indicated by the higher beta value.

5) Calculate the confidence intervals for the parameters in your model and explain what the results indicate.

```
cor.test(housing_data$Sale.Price, housing_data$year_built)
```

```
##
##  Pearson's product-moment correlation
##
## data:  housing_data$Sale.Price and housing_data$year_built
## t = 28.371, df = 12863, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2263401 0.2588660
## sample estimates:
##       cor
## 0.2426713
```

```
cor.test(housing_data$Sale.Price, housing_data$sq_ft_lot)
```

```
##
##  Pearson's product-moment correlation
##
## data:  housing_data$Sale.Price and housing_data$sq_ft_lot
## t = 13.687, df = 12863, p-value < 0.00000000000000022
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1027447 0.1368093
## sample estimates:
##       cor
## 0.1198122
```

The confidence interval for Sale.Price and year_built is 0.23-0.25 which gives us confidence that there is a positively linear relationship between the two variables.

The confidence interval for Sale.Price and sq_ft_lot is 0.10-0.13 which also gives us confidence that there is a positively linear relationship between the two variables.

Both variables give very small ranges of confidence interval which is good as it means our estimate is very likely correct as the correlation value is predicted to fall within those ranges 95% of the time.

6) Assess the improvement of the new model compared to your original model (simple regression model) by testing whether this change is significant by performing an analysis of variance.

```
anova(PriceBySqFtLot.lm, PriceBuildLotSize.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Sale.Price ~ sq_ft_lot
## Model 2: Sale.Price ~ sq_ft_lot + year_built
##   Res.Df             RSS Df       Sum of Sq      F              Pr(>F)
## 1  12863 2073376756946868
## 2  12862 1929833267833791  1 143543489113076 956.69 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F value is 956.69 and the p value is $<.001$ meaning that there is a significant improvement to the model by adding the year_built variable.

7) Perform casewise diagnostics to identify outliers and/or influential cases, storing each function's output in a dataframe assigned to a unique variable name.

```
housing_data$Residuals <- resid(PriceBuildLotSize.lm)
housing_data$StandResiduals <- rstandard(PriceBuildLotSize.lm)
housing_data$StudentResiduals <- rstudent(PriceBuildLotSize.lm)

housing_data$Cooks <- cooks.distance(PriceBuildLotSize.lm)
housing_data$DFBeta <- dfbeta(PriceBuildLotSize.lm)
housing_data$DFFit <- dffits(PriceBuildLotSize.lm)
housing_data$leverage <- hatvalues(PriceBuildLotSize.lm)
housing_data$covarRatio <- covratio(PriceBuildLotSize.lm)
```

8) Calculate the standardized residuals using the appropriate command, specifying those that are +-2, storing the results of large residuals in a variable you create.

```
housing_data$LargeResidual <- housing_data$StandResiduals > 2 | housing_data$StandResiduals < -2
```

9) Use the appropriate function to show the sum of large residuals.

```
sum(housing_data$LargeResidual)
```

```
## [1] 365
```

There are 365 large residuals out of the 12865 rows of data which is equal to roughly 3% of the data.

10) Which specific variables have large residuals (only cases that evaluate as TRUE)?

```
## # A tibble: 365 x 3
##    Sale.Price sq_ft_lot StandResiduals
##         <dbl>     <dbl>          <dbl>
## 1      165000    278891          -2.30
## 2     1445000     36446           2.02
## 3     1900000     37017           3.21
## 4     1520000     19173           2.88
## 5     1588359      8752           2.24
## 6     1450000     14043           2.40
## 7     1450000     14043           3.26
## 8      200000    288367          -2.19
## 9      275000    532739          -2.78
## 10      90000    574992          -3.22
## # ... with 355 more rows
```

11) Investigate further by calculating the leverage, cooks distance, and covariance ratios. Comment on all cases that are problematics.

```
## # A tibble: 365 x 3
##        Cooks   leverage covarRatio
##        <dbl>      <dbl>      <dbl>
## 1  0.00331   0.00187        1.00
## 2  0.000113  0.0000831      0.999
## 3  0.000289  0.0000845      0.998
## 4  0.00147   0.000530       0.999
## 5  0.000196  0.000117       0.999
## 6  0.000385  0.000201       0.999
## 7  0.00567   0.00159        0.999
## 8  0.00313   0.00195        1.00
## 9  0.0176    0.00677        1.01
## 10 0.0268    0.00769        1.01
## # ... with 355 more rows
```

Of the properties with large residuals:

- 2 of the properties come close to a Cook's distance of 1 (0.95 and 0.89) as the highest distance is 0.14.

- 19 properties have leverage values 2 times the average leverage value of 0.0002.

- 26 properties have a CVR value greater than $1.0004 = (1 + [3(2 + 1) / 12856])$ and 19 have a CVR value less than $0.9996 = (1 - [3(2 + 1) / 12856])$.

12) Perform the necessary calculations to assess the assumption of independence and state if the condition is met or not.

```
car::durbinWatsonTest(PriceBuildLotSize.lm)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.6203378     0.7593244       0
##  Alternative hypothesis: rho != 0
```

The Durbin Watson test returns a value of 0.76 which is not optimal as it is less than one. This indicates that the assumption of independence is not met in this data.

13) Perform the necessary calculations to assess the assumption of no multicollinearity and state if the condition is met or not.

```
library(car)
```

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

```
vif(PriceBuildLotSize.lm)
```

```
##  sq_ft_lot year_built
##   1.018539   1.018539
```

```
1 / vif(PriceBuildLotSize.lm)
```

```
##  sq_ft_lot year_built
##  0.9817982  0.9817982
```

```
mean(vif(PriceBuildLotSize.lm))
```
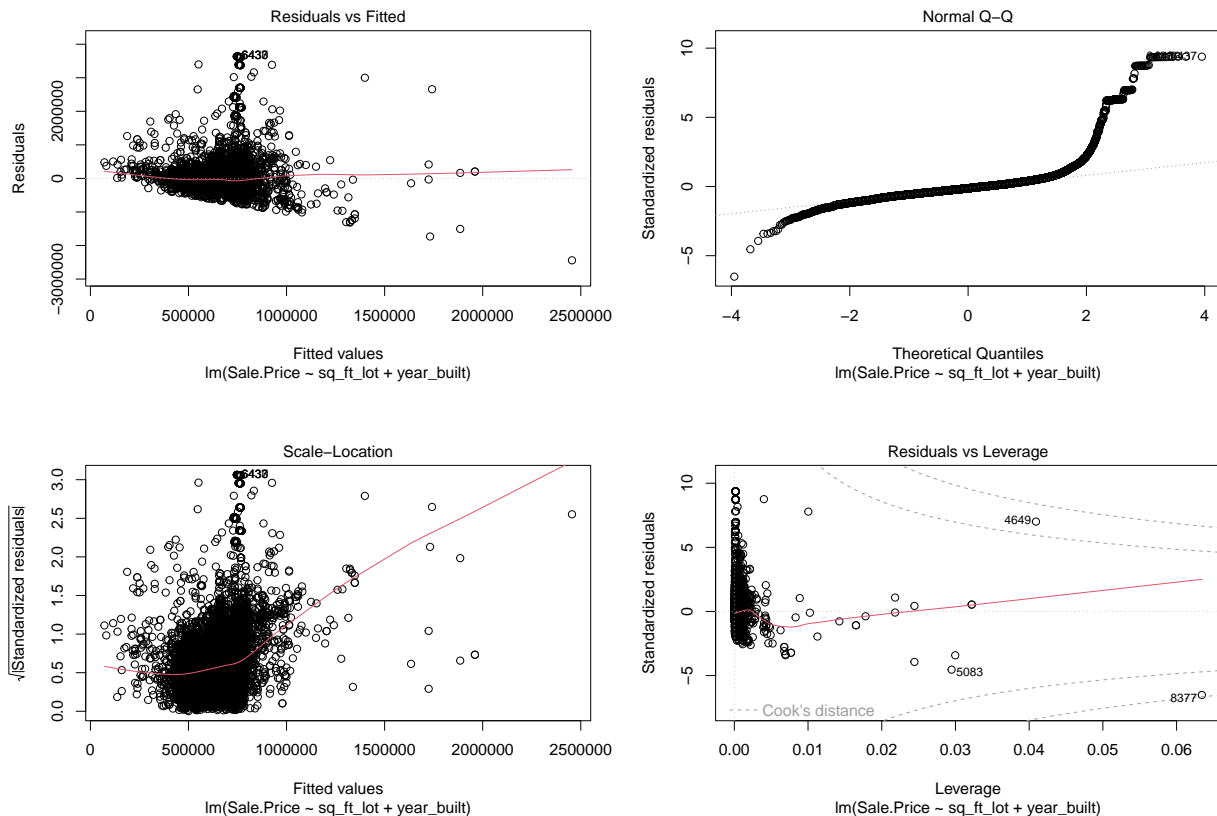
```
## [1] 1.018539
```

- VIF is not greater than 10 so no worries there.
- The average VIF is very slightly greater than 1 thus the regression may be every so slightly biased.
- Tolerance levels are around 0.98 so no worries there either.
- Based on these observations, we can safely conclude that there is no collinearity within the data.

14) Visually check the assumptions related to the residuals using the plot() and hist() functions. Summarize what each graph is informing you of and if any anomalies are present.
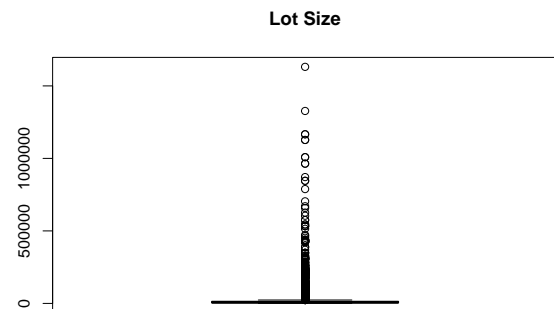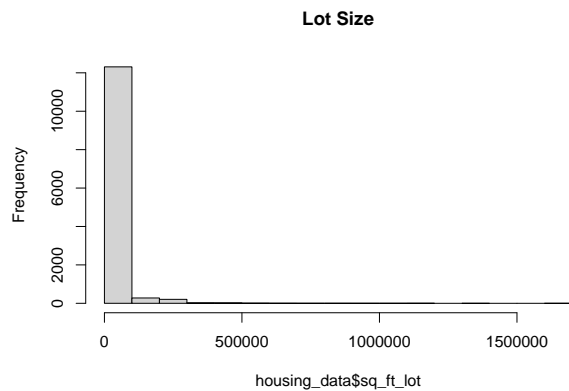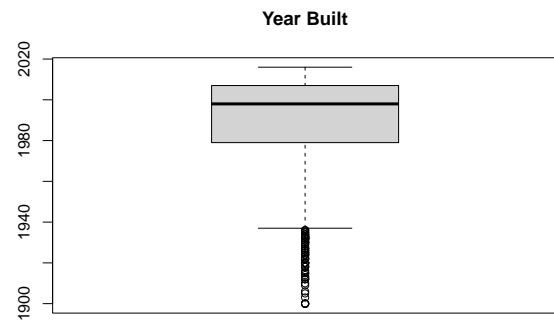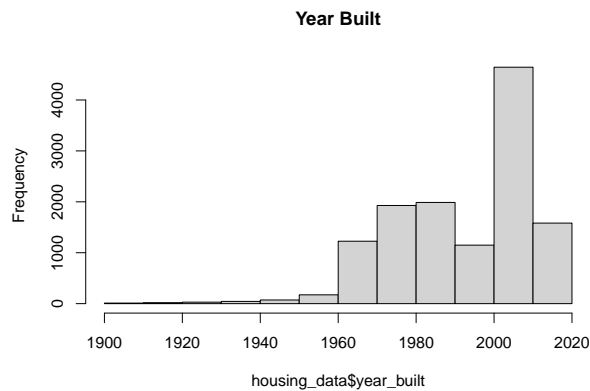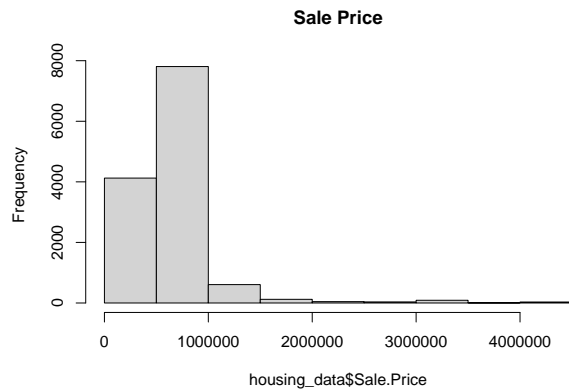
Sale.Price and sq_ft_lot both have positively skewed distributions while year_built has a negatively skewed distribution.The sq_ft_lot has the strongest skew.

The skewing shown in the histograms is echoed in boxplots of each of the variables showing that there are many potential outliers at the tail ends of the data.

Residuals vs Fitted

Residuals

Fitted values
lm(Sale.Price ~ sq_ft_lot + year_built)

Normal Q–Q

Standardized residuals

Theoretical Quantiles
lm(Sale.Price ~ sq_ft_lot + year_built)

Scale–Location

$\sqrt{|\text{Standardized residuals}|}$

Fitted values
lm(Sale.Price ~ sq_ft_lot + year_built)

Residuals vs Leverage

Standardized residuals

4649
5083
8377
Cook's distance

Leverage
lm(Sale.Price ~ sq_ft_lot + year_built)

The first plot the Residuals vs Fitted, shows a not so random pattern indicating that the assumptions of heteroscedasticity, linearity and randomness have NOT been met. This reinforces what we found earlier that the assumption of independence was NOT met in this data.

The second plot (The QQ Plot) shows that the values in the lower range do have some properties of normal distribution (based on the closeness of the data to line), but as the data value increases, the normal distribution disappears and the data points very farther from the line indicating a positive skew.

In the third plot, Scale-Location, the red line is not horizontal indicating the assumption of homoscedasticity is NOT met. Also the points are bunched up in the lower x values indicating greater variance at that end of the regression.

The last Residuals vs Leverage plot show there there is one data point (#8377) that has significant leverage (compared to the other data points in this set) but very little residual meaning it may have swayed the model to fit itself. Likewise towards the top of the graph point #4649 also has a high cook's distance, but has less leverage which resulted in a higher residual value as it didn't have as much sway as point 8377. Both points are influential points. Most of the data points have very high residuals meaning that they don't fit the linear model very well, but have little leverage to pull the line towards them.

15) Overall, is this regression model unbiased? If an unbiased regression model, what does this tell us about the sample vs. the entire population model?

- The VIF value was around 1 which indicates that there is very little multicollinearity bias present.

- If we consider points 8377 and 4649 outtliers, then they are applying some bias to the model, by pulling the linear regression towards themselves.

- Because the data appears to be unbiased, that means that we can assume that it is representative of the population.

Final Project Step 1

Introduction

Birdwatching is a very popular hobby in the United States and has a significant impact to the economy. Birdwatchers (or birders) spend money not only on their birdwatching gear, but often also travel to see specific species and thus promote ecotourism. With global warming, weather patterns are shifting with seasonality changes and temperature increases along with increased severity of weather events such as hurricanes. This changes could both impact bird populations and the eagerness of birders to go birdwatching.

I would like to look at how different weather aspects such as precipitation, temperature, and hurricanes effect bird observations both in terms of the number of checklists that are being submitted by birders and the diversity of species being observed across the state of Florida.

Data science allows us to bring together different information sources and datasets to look for patterns and hopefully provide insight into the larger picture than any one dataset alone. It also provides the rigor of statistical analysis to help us determine whether the patterns we perceive are merely coincidences or something more organized.

Research Questions

Temperature

1) What are the overall temperature trends?
2) Does temperature effect on the number of checklists submitted to eBird by birders?
3) Does temperature effect on the number of species observed by birders?

Precipitation

4) What are the overall precipitation trends?
5) Does precipitation effect on the number of checklists submitted to eBird by birders?
6) Does precipitation effect on the number of species observed by birders?

Hurricanes

7) What are the overall trends for number of hurricanes in a given year?
8) What are the overall trends for hurricanes intensity through the years?
9) Does the number of hurricanes in a given year effect the number of checklists submitted to eBird by birders?
10) Does the number of hurricanes in a given year effect on the number of species observed by birders?
11) Does average hurricane intensity in a given year effect on the number of species observed by birders?

Approach

My idea is to attempt to plot a series of linear regression models to try to ascertain trends within each of the variables and within variable interactions and see if any of the interactions are statistically significant. I anticipate that by using yearly averages of each variable that I will be able to avoid seasonal variances in the data both for the weather data and for the bird spring and fall migrations spiking bird species in the area.

Data

Bird Data

My primary dataset is from eBird.

eBird Basic Dataset. Version: EBD_relMay-2022. Cornell Lab of Ornithology, Ithaca, New York. May 2022. https://ebird.org/data/download

eBird is an online citizen science project started in 2002 by the Cornell Lab of Ornithology promoting conservation, wildlife education, and science-based studies. Birders from all over the world submit their checklists of sightings stating what species they saw and where and when they saw them. The goal of eBird data is to help scientists better track bird population trends both spatially and temporally. By involving the public, eBird is able to not only inspire a love of nature and sense of involvement, but at the same time collect an immense amount of data to see global bird population trends. Data entry from the public is reviewed by regional volunteers to promote consistency and accuracy.

The original data set has 49 columns. I requested all data from Florida only, but the dataset can be customized for a particular location and time period. Much of the variables are blank due to lack of input from the original birder during submission. These variables will not be useful for study, but certain fields like the date, species, and location among others are required fields and thus present. In the number observed column, it is possible to encounter an X where the birder submitted that the species was present, but the birder did not count how many individuals of that species there were. The data also includes a unique sampling event identifier which enables us to group observations by checklists.

Weather Data

My other two datasets are both come from the Florida Climate Center. The Florida Climate Center provides historical weather data and related information to any interested party and collaborates with the National Climatic Data Center. The datasets I am using both show the respective weather statistic (precipitation and temperature) for each month and the yearly average ranging from as far back as January 1895 to June 2022.

Temperature: https://climatecenter.fsu.edu/products-services/data/statewide-averages/temperature

Precipitation: https://climatecenter.fsu.edu/products-services/data/statewide-averages/precipitation

Since the eBird data only goes back to 2002, I won't technically need the weather data before that date, but It would be helpful when looking separately at the weather trends.

Required Packages

I will need/want:

- readxl – for reading excel files
- ggplot2 – for graphing
- auk – for working with eBird data

- ppcor – for partial correlations between variables

And also to wrangle the data:

- formattable
- plyr
- dplyr
- purrr

And probably others that I will discover I need as I go along and try to do something.

Plots and Table Needs

At very least I will want to graph each variable over time to see the general trends. Then I will also want to see if each variable has a normal distribution of values. I will also want to plot the interactions of each of the variables to see if a linear regression would fit the data well or not.

Questions for future steps

Since I haven't yet started plotting the data, I am not sure exactly what I will need. I could foresee the data not fitting a linear trend and therefore might need to learn how to work with non-linear modeling.

The eBird data is in a .txt file that is larger than excel can handle, so I am going to have to figure out how to get R to read that directly, but I think that should be similar to reading a csv file.

There are definitely things I need to get stronger on, but honestly I still feel like I am at the point where I don't' know what I don't know. I feel like I will find out as I go to do something and then realize I don't yet know how to do it.