308_Final_Project

Amulya Bollapragada

2025-04-21

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

Part 1

Question 1

R markdown is a useful tool because it makes the coding process more reproducible. This is because a markdown file allows us to write down our thought process and steps more easily while coding. With R markdown, we can create a document with text of explanation, code chunks, as well as the outputs. It is also a helpful tool for us to collaborate and communicate with fellow statisticians and other professionals. It is a very powerful platform for data science!

Question 2

Read the house dataset into R and save it as an R object. Then, display the first 6 rows of the dataset:

```
house <- read_csv("aacline_house.csv")
```

```
## Rows: 1000 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): HouseStyle, BsmtFinType2, Foundation, RoofStyle, Exterior1st, LandS...
## dbl (7): SalePrice, BsmtFinSF1, GarageCars, YearBuilt, OpenPorchSF, WoodDeck...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
slice(house, 1:6)
```

```
## # A tibble: 6 x 13
##   SalePrice BsmtFinSF1 GarageCars YearBuilt OpenPorchSF WoodDeckSF YearRemodAdd
##       <dbl>      <dbl>      <dbl>     <dbl>       <dbl>      <dbl>        <dbl>
## 1    223500        486          2      2001          42          0         2002
## 2    140000        216          3      1915          35          0         1970
## 3    250000        655          3      2000          84        192         2000
## 4    200000        859          2      1973         204        235         1973
## 5    129500        906          1      1965           0          0         1965
## 6    345000        998          3      2005          21        147         2006
## # i 6 more variables: HouseStyle <chr>, BsmtFinType2 <chr>, Foundation <chr>,
## #   RoofStyle <chr>, Exterior1st <chr>, LandSlope <chr>
```

Question 3

The GarageCars variable has under it which indicates that it is a numeric value. Thus, it is a quantitative variable.

The HouseStyle variable has a under it which indicates that it is a character value. Thus, it is a categorical variable.

Question 4

Display a tibble that only contains homes that are built after the median value of YearBuilt, have a sales price lower than the mean value of SalesPrice, and are classified as 2Story homes:

```r
house %>%
  filter(YearBuilt > median(YearBuilt)) %>%
  filter(SalePrice < mean(SalePrice)) %>%
  filter(HouseStyle == '2Story')
```

```
## # A tibble: 124 x 13
##    SalePrice BsmtFinSF1 GarageCars YearBuilt OpenPorchSF WoodDeckSF YearRemodAdd
##        <dbl>      <dbl>      <dbl>     <dbl>       <dbl>      <dbl>        <dbl>
##  1    223500        486          2      2001          42          0         2002
##  2    172500        649          2      1999           0        115         2000
##  3    196500          0          2      2004          70          0         2004
##  4    185000          0          2      1998          94          0         1998
##  5    174000          0          2      2005          38        100         2005
##  6    204750        648          2      1997         162          0         1997
##  7    178000          0          2      1985          46        192         1985
##  8    176000        419          2      1999          32          0         1999
##  9    163990          0          2      2005          24          0         2006
## 10    130000          0          2      2004          40          0         2006
## # i 114 more rows
## # i 6 more variables: HouseStyle <chr>, BsmtFinType2 <chr>, Foundation <chr>,
## #   RoofStyle <chr>, Exterior1st <chr>, LandSlope <chr>
```

Question 5

Make the following modifications to your house dataset: – rename SalePrice to sale_price – rename GarageCars to garage_cars – rename YearBuilt to year_built – Change the variable type of garage_cars to be a factor – Ensure that this data set only contains these three columns.

Next, print the first 6 rows of this data set.

```
house_subset <- house %>%
  rename("sale_price" = "SalePrice",
         "garage_cars" = "GarageCars",
         "year_built" = "YearBuilt") %>%
  mutate("garage_cars" = factor(garage_cars, levels = c(1,2,3,4))) %>%
  select("sale_price", "garage_cars", "year_built")

head(house_subset)
```

```
## # A tibble: 6 x 3
##   sale_price garage_cars year_built
##        <dbl> <fct>            <dbl>
## 1     223500 2                 2001
## 2     140000 3                 1915
## 3     250000 3                 2000
## 4     200000 2                 1973
## 5     129500 1                 1965
## 6     345000 3                 2005
```

Question 6

Write a function that converts sales price from US to Canadian dollars. Then, run the function on your data and print out the first 5 values:

```
exchange <- function(US_dollars) {
    CA_dollars = US_dollars*1.45
    return(CA_dollars)
}

exchange(house$SalePrice)[1:5]
```

```
## [1] 324075 203000 362500 290000 187775
```

Question 7

Recreate a sample plot for the house subset data:

```
g <- ggplot(house_subset, aes(x = year_built, y = sale_price), warning = false)
g + geom_point(aes(color = garage_cars)) +
geom_smooth(aes(color = garage_cars), se = F) +
labs( x = "Years", y = "Sales Price", title = "Year Built vs Sales Price by Garage", color = "Garage")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

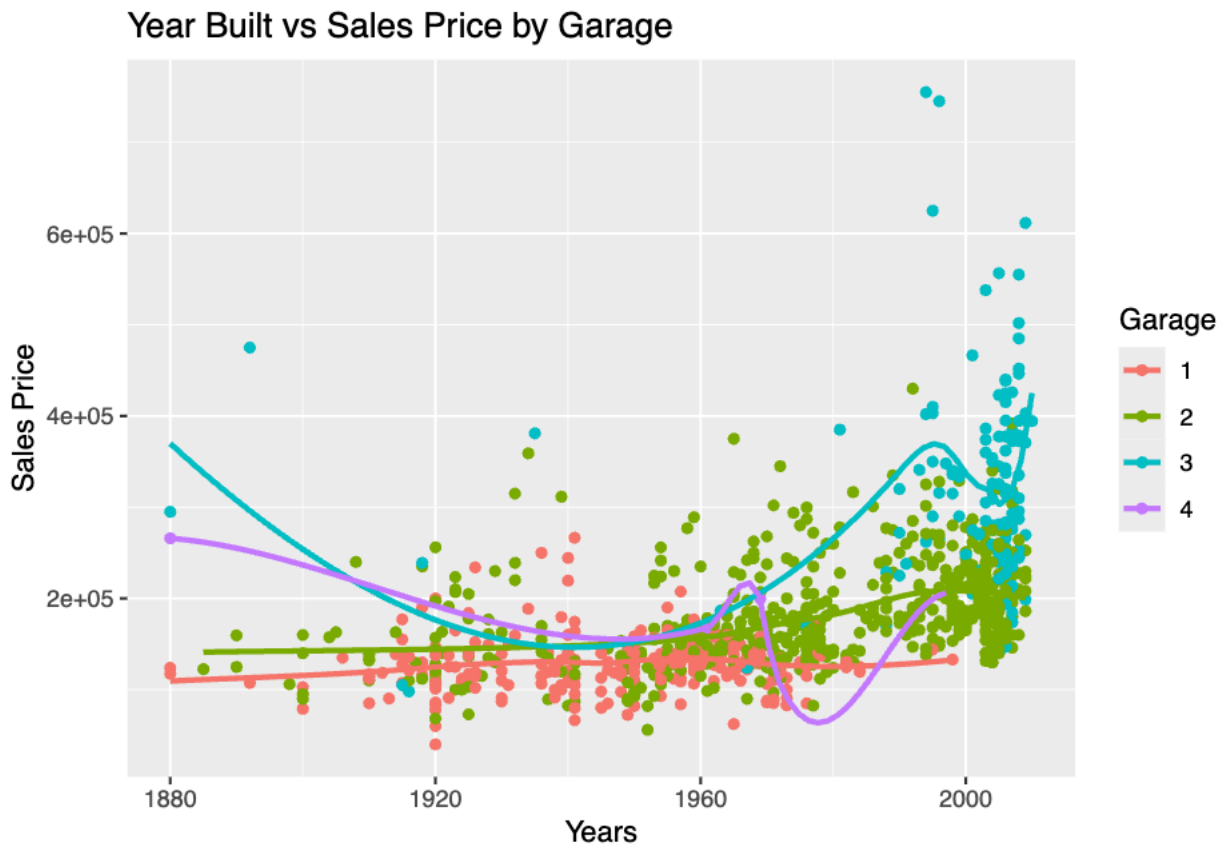
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : span too small.  fewer data values than degrees of freedom.

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 1879.4

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 89.585

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 817.1
```



According to the plot, sales price for cars in garages 2 and 3 generally increased as years pass. The sales price for cars in garage 1 stayed about the same as years passed. The sales price for cars in garage 4 had many fluctuations over the years.

Part 2

Question 1

I chose to look at a Math Students Performance dataset.

I got this dataset at this link: https://www.kaggle.com/datasets/adilshamim8/math-students

This dataset includes grades for each student. Along with their grades, the dataset also includes information about the rest of their life such as family size, parents' education and jobs, desire to pursue higher education, and much more. The full list of these variables and their descriptions can be found on the website link above. According to the website that this dataset is from "This dataset originates from the UCI Machine Learning Repository and was originally featured in the study by P. Cortez and A. Silva, titled Using Data Mining to Predict Secondary School Student Performance. The dataset was collected as part of research presented at the 5th FUBUTEC 2008 conference in Porto, Portugal".My research question is "Do students with at home internet access have higher final grades than those who do not?". I think this question is interesting because regardless of how much effort a teacher can put in, average grades can still be low for a variety of reasons. Instead of directly placing blame on instructors, it can be a good idea to investigate what other variables impact a student's final grade. After knowing this, maybe school faculty can help students ensure that certain areas of their life don't negatively impact their grades. My research question only makes a small step to investigating for this cause. Conducting inference and research on how/if all or any of these variables impact grades can be very helpful for school faculty to have a comprehensive understanding of what may be impacting their students' grades.

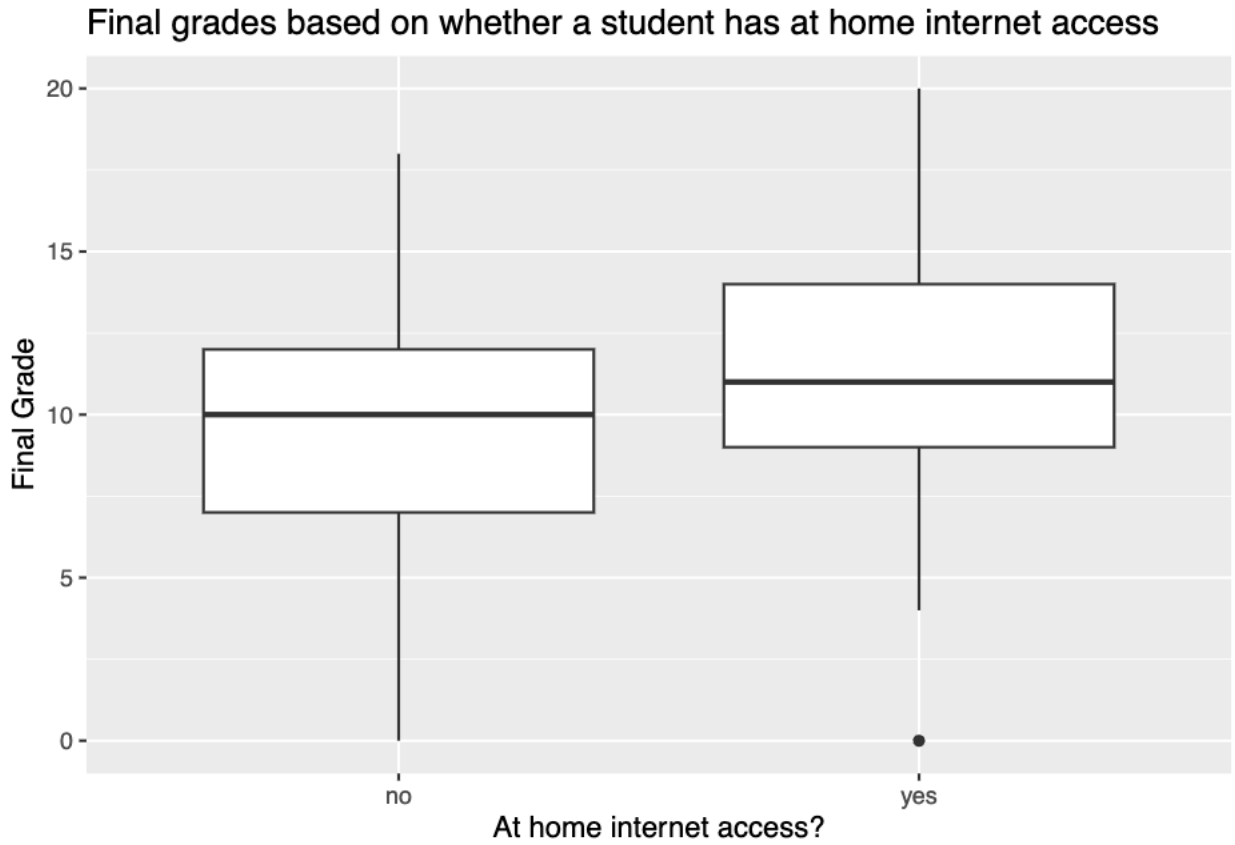
Question 2

To investigate my research question, I chose to make a box plot of the final grade distributions subsetting by whether or not a student has internet access.

```
math_grades <- read_csv("Math-Students.csv")
```

```
## Rows: 399 Columns: 33
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (17): school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardi...
## dbl (16): age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
g <- ggplot(math_grades, aes(x= internet, y = G3))
g + geom_boxplot() +
labs(title = "Final grades based on whether a student has at home internet access", x = "At home intern
```



According to my box plot, the final grade distribution for students with at home internet access is slightly higher than that of students without at home internet access. The maximum, minimum, and median grades of students with at home internet access are higher than the that of students without at home internet access. Thus, there is evidence for the claim that students with at home internet access have higher final grades than students who do not. Using this information, education systems can try to provide more assistance to students without at home internet access to hopefully improve their grades.