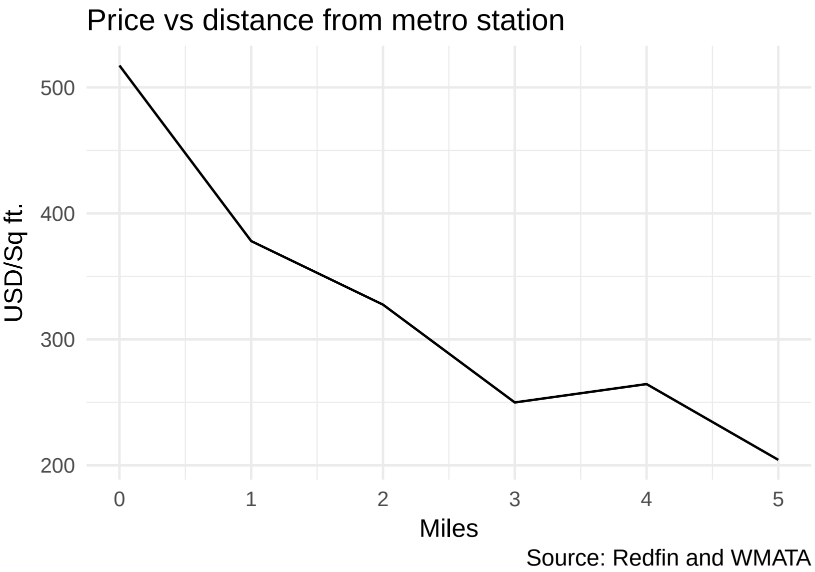
# Module 2

# Day 4

# Recap Last Week

- Discussed long and wide data
  - Reshaped the data using `spread()` and `gather()`
- Created our own set of functions to calculate distance to the nearest metro
  - Wrapper functions
- Discussed function safety
  - `require()`, `warning()`, `stop()`

# Recap Last Week



## Goals for Today

R:

- Review liner models and dates
- Create non-linear models in R

Economics:

- Further exploring the relationship between distance from a metro stop and home prices
- Investigate non-linear relationship of location, distance, time on the market on home prices
- Familiarize ourselves with different methods to account for non-linear effects

# Brief Review of Regressions

- Use it when we want to indentify a causal relationship
- Regression analysis is used to describe the relationship between:
  - A single response variable $Y$ and
  - One or more predictor variables $X_1, X_2, X_3, \ldots, X_n$
- Response should be continuous (but doesn't have to be!)
- Predictor variables can be continous, discrete, or categorical
  - Dummy variables are used to model categorical data

# Creating an OLS Model

- Let's load in the data and see if there is a causal relationship between distance from the metro and home prices

```
library(tidyverse)

setwd( ) # Put your  file directory here
joined_data <- read_csv("joined_data_v2.csv")
```

- What function do we use to create a regression model?
- What function do we use to create our regression tables?

# The Effect of Metro Distance on Home Prices

```
library(stargazer)

dist_OLS <- lm(PRICE/SQUARE.FEET ~ metro_distance, data = joined_data)
stargazer(dist_OLS, header = F, dep.var.caption = "",
          title = "Impact of Distance From Metro on Price",
          omit.stat = c("ser", "f"),
          no.space = T)
```

Table 1: Impact of Distance From Metro on Price

| | PRICE/SQUARE.FEET |
|---|---|
| metro_distance | $-93.621^{***}$ |
| | (6.467) |
| Constant | $503.684^{***}$ |
| | (8.202) |
| Observations | 1,179 |
| $R^2$ | 0.151 |
| Adjusted $R^2$ | 0.150 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

# Interpreting the Results

- How do we interpret the coefficient for `metro_distance`?
- Does our intercept have any meaningful interpretation?
- Is our current model a good model?
  - If not, how could we improve it?

# Adding States to the Model

- What are factors that affect housing prices that could differ from state to state?

```
state_dist <- lm(PRICE/SQUARE.FEET ~ metro_distance +
                                     STATE,
                data = joined_data)
```

# Comparing the Results

Table 2: Impact of Distance From Metro on Price

| | PRICE/SQUARE.FEET | |
|---|---|---|
| | (1) | (2) |
| metro_distance | $-93.621^{***}$ | $-81.569^{***}$ |
| | (6.467) | (7.352) |
| STATEMD | | $-54.973^{***}$ |
| | | (14.816) |
| STATEVA | | 0.995 |
| | | (13.996) |
| Constant | $503.684^{***}$ | $505.487^{***}$ |
| | (8.202) | (8.592) |
| Observations | 1,179 | 1,179 |
| $R^2$ | 0.151 | 0.163 |
| Adjusted $R^2$ | 0.150 | 0.161 |

# Comparing the Results

- How do we interpret the coefficients on our dummy variables?
- What happened to our `metro_distance` coefficient when we added our dummy variables?

# Days on the Market

- Conventional wisdom says that homes that spend a long time on the market generally cost less
  - Seller may have set the price to high originally
  - Buyers may be waiting for the seller to lower the price
- Our data has the list date and the sale date
  - Work with dates to create a new variable `days_on_market`

# Dates Review

- In R, dates are just numbers displayed in a special format
  - Generally stored as strings in our data
  - Use `as.Date()` to convert them from strings to date objects

```r
date1 <- as.Date("March-26-2018", format = "%B-%d-%Y")
date2 <- as.Date("Wednesday- Mar 28, 2018",
                 format = "%A- %b %d, %Y")
```

- We are able to read dates in virtually any format that they are entered
  - We can find a list of formats here

# Dates Review

- Since dates are just numbers we can use them with mathematical functions and operations

```
date2 - date1
```

```
## Time difference of 2 days
```

```
date1 - 3
```

```
## [1] "2018-03-23"
```

- We can also generate date sequences

```
seq(date1, date2, by = "day")
```

```
## [1] "2018-03-26" "2018-03-27" "2018-03-28"
```

# Refresher Exercise

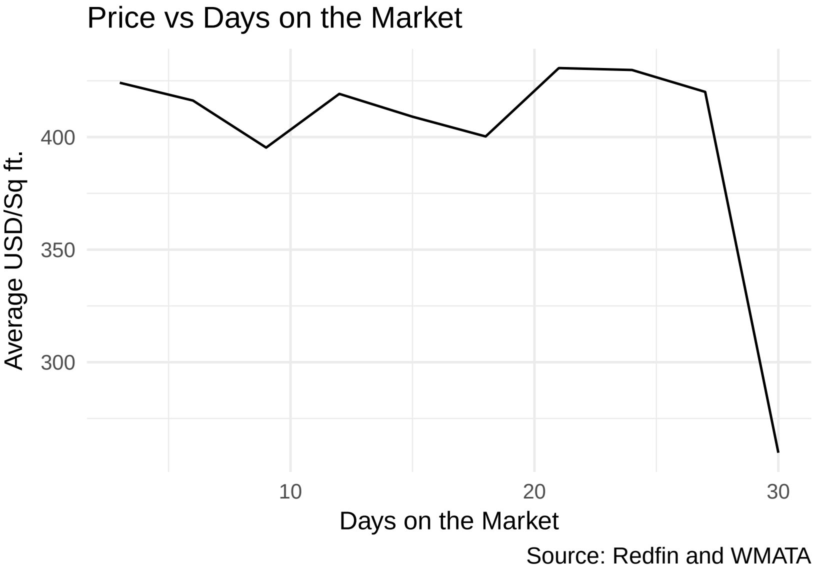
- Load in the dates.csv data file then:
  - Combine the three variables `Year`, `Month`, and `Day` into a single string variable `Date`
  - Convert `Date` into a date variable

# Creating days_on_market

- Remember dates are just numbers!
- We can calculate the number of days on the market by merely subtracting the sale date by the list date and converting it back to a number

```
joined_data <- joined_data %>%
    mutate(SOLD.DATE = as.Date(SOLD.DATE, format = "%B-%d-%Y"),
           LIST.DATE = as.Date(LIST.DATE, format = "%A- %b %d, %Y"),
           days_on_market = as.numeric(SOLD.DATE - LIST.DATE))
```

# Graphing the Relationship



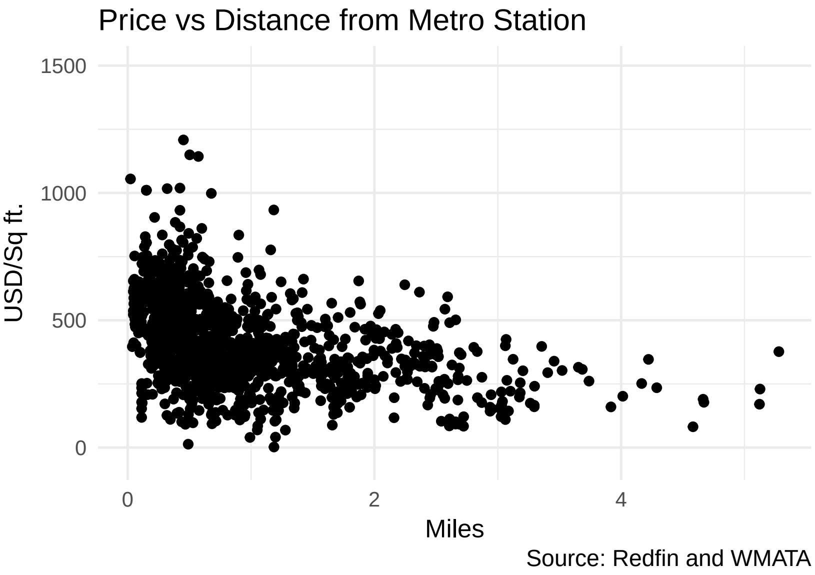
# Testing Our Hypothesis

- The relationship doesn't appear to be as straightforward as we initially thought
- Create a regression model that investigates the relationship between the days a house spends on the market and the price per square foot of a home
  - You are free to use any variable in our data set
  - Use `stargazer()` to display your results
  - Why do you believe your model is a good model?

# Introduction to Non-Linear Models

- Made some improvements to our model, but we can still do better
- Let's think more about metro distance
  - As we move farther away from the metro, distance probably matters less
  - The effect of distance likely varies from state to state
- We are unable to investigate either of the above relationships with a simple linear model

# Looking Back at Distance



# Adding A Quadratic Term

- In our previous models, moving from 0.5 to 0.6 miles away from the metro has the same effect on price as 2.9 to 3.0 miles away
- Can imagine that people pay extra to live in walking distance
  - The farther we move away from the metro, the less distance affects price
- We are able to apply mathematical transformations to the continous variables in our model
  - We would like to use a quadratic term in our model

# Adding A Quadratic Term

- There are two ways we can include a quadractic term in our data:
  1. Using formula functions provided by R
  2. Constructing a new variable called `dist_sq` and adding it to the model
- First, let's try to use the caret to take the exponenet

```
dist_sq <- lm(PRICE/SQUARE.FEET ~ metro_distance +
                                  metro_distance^2,
                data = joined_data)
```

# Did It Work?

Table 3: Impact of Distance From Metro on Price

| | PRICE/SQUARE.FEET | |
|---|---|---|
| | (1) | (2) |
| metro_distance | $-93.621^{***}$ | $-93.621^{***}$ |
| | (6.467) | (6.467) |
| Constant | $503.684^{***}$ | $503.684^{***}$ |
| | (8.202) | (8.202) |
| Observations | 1,179 | 1,179 |
| $R^2$ | 0.151 | 0.151 |
| Adjusted $R^2$ | 0.150 | 0.150 |
| *Note:* | $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ | |

# Correcting Our Formula

- When we are working with formulas in R, our math operators have different meanings
  - ^ and * are used to create interactions
- Use `I()` to apply our math operators

```
dist_sq <- lm(PRICE/SQUARE.FEET ~ metro_distance +
                                  I(metro_distance^2),
                data = joined_data)
```

# Did It Work?

Table 4: Impact of Distance From Metro on Price

| | PRICE/SQUARE.FEET | |
|---|---|---|
| | (1) | (2) |
| metro_distance | $-93.621^{***}$ | $-200.919^{***}$ |
| | (6.467) | (17.263) |
| I(metro_distance^2) | | $32.547^{***}$ |
| | | (4.869) |
| Constant | $503.684^{***}$ | $554.116^{***}$ |
| | (8.202) | (11.036) |
| Observations | 1,179 | 1,179 |
| $R^2$ | 0.151 | 0.182 |
| Adjusted $R^2$ | 0.150 | 0.181 |
| *Note:* | $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$ | |

# Manually Constructing the Variable

- We should be right but let's double check to be sure
- Create a new variable `metro_sq` and use it to create a non-linear regression model
  - Compare this new model with the previous model in `stargazer()`
  - Are the coefficients the same?

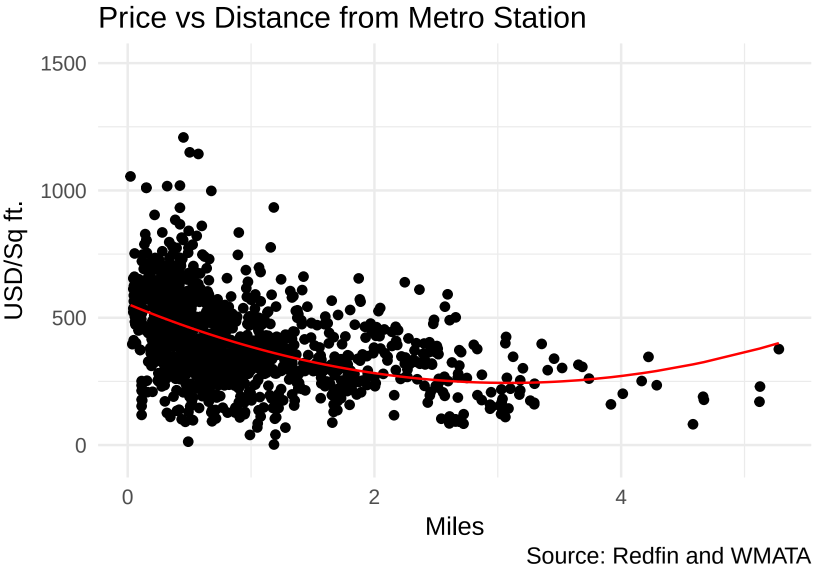
# Interpreting Our Results

- Let's look back at our regression table, do the magnitidues in our table make sense?
- Direct interpretation of our coefficients becomes much more difficult
  - The effect changes as we move farther away from a metro station
  - Should check when the effect of metro distance becomes 0
  - Can evaluate the effect at the mean metro distance
- Helpful to create a graph to visualize this relationship

# broom()

- Recall our 3 main functions from the broom package
  - `tidy()` - for creating a data frame of component statistics
  - `augment()` - for observation level statistics (like fitted values and residuals)
  - `glance()`- for model level statistics (like R-squared etc.)
- We'll want to use `augment()` to plot our fitted values

# Visulaizing the Relationship



# Visulaizing the Relationship

- Fill in the code to create the graph from the previous slide

```
library(broom)
dist_sq_augmented <- augment( , )

dist_sq_augmented %>%
  mutate(price = PRICE/SQUARE.FEET) %>%
ggplot() +
  geom_point(aes(x = , y = ))+
  geom_line(aes(x = , y = ), color = ) +
  labs(title = "Price vs Distance from Metro Station",
       y = "USD/Sq ft.", x = "Miles",
       caption = "Source: Redfin and WMATA")+
  scale_y_continuous(limits = c(-50, 1500)) +
  theme_minimal()
```

# Interactions

- Sometimes the effect of certain variables differs across groups in our data
- The importance of distance from the metro likely varies from state to state
  - Why might that be the case?
- To account for this type of non-linear we use an interaction term
- Below is an model expressed as an equation:
  - $\text{Price/Sq Ft} = \beta_0 + \beta_1 \text{distance} + \beta_2 \text{MD} + \beta_3 \text{VA} + \beta_4 \text{MD} \times \text{distance} + \beta_5 \text{VA} \times \text{distance}$