Video 2.1 A Customer Analytics Dataset to Illustrate Indicator Variables
The instructor used a simulated data set which mimics data from a direct marketing firm. The data set contains variables including indicator variables and numerical variables. **Indicator variable is also known as dummy variable**. We are trying to predict the amount customers spent on buying products using customer characteristics such as age (indicator variable), salary (numerical variable), location (indicator variable). A quick look of the dataset:

First 10 Rows of the *dirmkt* Dataframe

Age	Gender	OwnHome	Married	Location	Salary	Children	History	Catalogs	AmountSpent
Old	Female	Own	Single	Far	47500	0	High	6	75.5
Middle	Male	Rent	Single	Close	63600	0	High	6	131.8
Young	Female	Rent	Single	Close	13500	0	Low	18	29.6
Middle	Male	Own	Married	Close	85600	1	High	18	243.6
Middle	Female	Own	Single	Close	68400	0	High	12	130.4
Young	Male	Own	Married	Close	30400	0	Low	6	49.5
Middle	Female	Rent	Single	Close	48100	0	Medium	12	78.2
Middle	Male	Own	Single	Close	68400	0	High	18	115.5
Middle	Female	Own	Married	Close	51900	3	Low	6	15.8
Old	Male	Own	Married	Far	80700	0	None	18	303.4

The instructor first ran a regression on whether salary has an influence on AmountSpent. The result is as follows:

	Estimate	S.E.	t Value	Pr>|t|
Intercept	-1.531783	4.537416	-0.338	0.736
Salary	0.002196	0.000071	30.930 ***	<.001

R-squared	Adjusted R-squared
0.722	0.721

Note that the adjusted R-squared is the R-squared value adjusted for degree of freedom.

Video 2.2 Creating and Using Indicator (Dummy) Variables
The instructor first tests whether categorical variable 'Age' has an effect on AmountSpent. 'Age' has three possible values: Young, Middle, or Old. To create indicator variables for Age, we need two indicator (dummy) variables. The base case, with both dummy variables set to 0, is Age = Young. It is up to modeler to determine which value of the categorical variable is used as the base case. Therefore, the two dummy variables we have are:

$$\text{AgeMid} = \begin{cases} 1, & if\ Age = Middle \\ 0, & otherwise \end{cases}$$

$$\text{AgeOld} = \begin{cases} 1, & if\ Age = Old \\ 0, & otherwise \end{cases}$$

For example, when AgeMid = 0, AgeOld = 1, the record is for someone whose age is old. Note that AgeMid and AgeOld cannot be 1 at the same time since every individual has to be in exactly one age category.

Video 2.3 Interpreting the Coefficients of Indicator Variables
The instructor runs the regression model

$$AmountSpent = b_0 + b_1 * AgeMid + b_2 * AgeOld$$

Here is the result,

	Estimate	S.E.	t Value	Pr>|t|
Intercept	55.862	5.112	10.93***	<.001
AgeMid	94.307	6.395	14.75***	<.001
AgeOld	87.350	7.919	11.03***	<.001

Note that with AgeMid and AgeOld are 0, b_0 captures the average AmountSpent of customers who are Young. With AgeMid = 1 and AgeOld = 0, $b_0 + b_1$ capture the average AmountSpent of customer who are middle-aged. $94.307 is the increase in AmountSpent (on average) for middle-aged customers compared to young customers.

In R, we can use a Factor Variable in regression to create dummy variables
lm(AmountSpent ~ Age, data = dirmkt)
where dirmkt is the dataset. R's indicator variable coding scheme can be found by using:
contrasts(dirmkt$Age)

Next the instructor runs 2nd regression with Salary and dummy variable Age.

$$AmountSpent = b_0 + b_1 * Salary + b_2 * AgeMid + b_3 * AgeOld$$

The result is as follows:

	Estimate	S.E.	t Value	Pr>|t|
Intercept	-6.12	4.72	-1.30	0.20
Salary	.002	.00009	25	<.001
AgeMid	-4.81	6.39	-0.75	0.45
AgeOld	23.28	6.72	3.46	<.001

For one unit increase in salary, the average AmountSpent increases by $0.002.

Here is a quiz in the video, I think the answer is straightforward:

- What does this result mean?

A. Middle-aged customers spend the most
B. Old customers spend the least
C. Old customers spend more that young customers
D. At the same salary level, old customers spend more than young customers

What is the current answer?

D. At the same salary level, old customers spend more than young customers

D is the correct answer.

Video 2.4 Interaction Term and Interpreting its Coefficient
The instructor first runs a regression using variables Salary and Location. Location is a categorical variable with ‘Close’ if the customer lives close to a store and ‘Far’ otherwise.

$$AmountSpent = b_0 + b_1 * Salary + b_2 * Far$$

The regression result is:

	Estimate	S.E.	t Value	Pr>|t|
Intercept	-20.480	4.413	-4.64	<.0001
Salary	0.002	0.00007	34.05	<.0001
Far	59.060	4.414	13.38	<.0001

Multiple R-Squared: 0.5672, Adjusted R-squared: 0.5663

However, in this above model, **we assume that customers who live far away from a store that sells similar products will spend at the same rate as customers who live close to a store.** For example, if the salary increases by $1000, no matter the customer lives far away or close, the average AmountSpent increase is $2. One way of extending this model to allow for interaction effects is to include a third preditor, called an **interaction term**. In this case, we add a new variable SalaryFar which is Salary * Far.

$$AmountSpent = b_0 + b_1 * Salary + b_2 * Far + b_3 * SalaryFar$$

The regression result is:

	Estimate	S.E.	t Value	Pr>|t|
Intercept	1.448	4.808	0.30	0.76
Salary	0.002	0.000	24.72	<.0001
Far	-13.460	8.680	-1.55	0.12
SalaryFar	0.001	0.000	9.57	<.0001

Multiple R-Squared: 0.6036, Adjusted R-squared: 0.6024

The coefficient b_3 is the amount to add to b_1 to get the slope for individuals who live far away.
If the salary of a customer who lives close increases by $10,000, the predicted increase in AmountSpent is 0.002 * $10000 = $20.
If the salary of a customer who lives far away increases by $10,000, the predicted increase in AmountSpent is (0.002 + 0.001) * $10000 = $30.

Video 2.5 Another Example of Using Indicator Variables

The instructor uses a concrete example to summarize what we have learnt. The dataset is an AirBnB for Los Angeles Rental Market.
Listing data on AirBnB is publicly available at http://insideairbnb.com/los-angeles/ and http://insideairbnb.com/get-the-data.html

About the data used:

- Listing data collected on May 2, 2017
- We discarded listings with price greater than $1000 and missing values for beds, baths, and rating

```
$ Price           : num  50 55 150 30 45 80 120 55 50 50 ...
$ Reviews         : int  33 14 22 3 38 42 15 58 19 1 ...
$ Beds            : int  1 1 3 1 1 2 1 2 1 1 ...
$ Baths           : num  1 1 1 1 1 1.5 1 2 0 2 ...
$ Capacity        : int  2 2 6 1 2 2 2 3 1 2 ...
$ Monthly_Reviews : num  1.91 1.72 2.12 0.18 7.92 1.89 1.96 2.98 0.53 0.04 ...
$ Room_Type       : Factor w/ 3 levels "Shared room",..: 2 2 3 2 2 2 3 2 2 2 ...
$ Rating          : int  93 100 100 93 98 99 99 92 89 NA ...
```

We are going to find whether there is a relationship between capacity and price and whether Room_Type (shared, private, or full house) changes this relationship.

Data wrangling

```
la_listing <-  la_listing %>%
        mutate(Price = str_replace(Price, "[$]", "")) %>%
        mutate(Price = str_replace(Price, "[,]", "")) %>%
        mutate(Price = as.numeric(Price)) %>%
        mutate(Room_Type = factor(Room_Type, levels = c("Shared room", "Private room", "Entire home/apt"))) %>%
        mutate(Capacity_Sqr = Capacity * Capacity) %>%
        mutate(Beds_Sqr = Beds * Beds) %>%
        mutate(Baths_Sqr = Baths * Baths) %>%
        mutate(ln_Reviews = log(1+Reviews)) %>%
        mutate(ln_Monthly_Reviews = log(1+Monthly_Reviews))
        mutate(ln_Price = log(1+Price)) %>%
        mutate(ln_Beds = log(1+Beds)) %>%
        mutate(ln_Baths = log(1+Baths)) %>%
        mutate(ln_Capacity = log(1+Capacity)) %>%
        mutate(ln_Rating = log(1+Rating)) %>%
        mutate(Shared_ind = ifelse(Room_Type == "Shared room",1,0)) %>%
        mutate(House_ind = ifelse(Room_Type == "Entire home/apt",1,0)) %>%
        mutate(Private_ind = ifelse(Room_Type == "Private room",1,0)) %>%
        mutate(Capacity_x_Shared_ind = Shared_ind * Capacity) %>%
        mutate(Capacity_x_House_ind = House_ind * Capacity) %>%
        mutate(Capacity_x_Private_ind = Private_ind * Capacity) %>%
        mutate(ln_Capacity_x_Shared_ind = Shared_ind * ln_Capacity) %>%
        mutate(ln_Capacity_x_House_ind = House_ind * ln_Capacity) %>%
        mutate(ln_Capacity_x_Private_ind = Private_ind * ln_Capacity)
      filter(Price < 1000 , !is.na(Beds), !is.na(Baths), !is.na(Price), !is.na(Rating))
```

Convert price to numeric and room_type to factor

Create squared terms for testing non-linear relations

Create log terms for testing non-linear relations

Create dummy variables for room_type

Create interaction terms

Filter unwanted data

First we run regression on Price against Capacity.

$$Price = b_0 + b_1 * Capacity$$

The result is

	Estimate	**S.E.**	**t Value**	**Pr>|t|**
Intercept	15.039	1.141	13.19***	<.001
Capacity	38.272	0.316	114.72***	<.001

R-squared	**Adjusted R-squared**
0.367	0.367

Next, we use Room_Type to run regression:

$$Price = b_0 + b_1 * Private\_ind + b_2 * House\_ind$$

Note that the base case is ‘Shared’. The result is

	Estimate	S.E.	t Value	Pr>|t|
Intercept	37.149	2.954	12.58***	<.001
Private_ind	35.666	3.123	11.42***	<.001
House-ind	133.442	3.058	43.64***	<.001

The shared room's average price is $37.149; the private room's average price is $37.149 + $ 35.666; the house's average price is $37.149 + $133.442.

Next we run regression using Capacity and Room_Type,

$$Price = b_0 + b_1 * Capacity + b_2 * Private\_ind + b_3 * House\_ind$$

The regression result is

	Estimate	S.E.	t Value	Pr>|t|
Intercept	-19.017	2.678	-7.101	<.001
Capacity	29.292	0.355	82.605	<.001
Private_ind	30.339	2.739	11.076	<.001
House-ind	75.776	2.771	27.346	<.001

The average price increase for each additional individual is $29.292.

Next we are adding interaction terms:

$$Price = b_0 + b_1 * Capacity + b_2 * Private\_ind + b_3 * House\_ind + b_4 * P\_Cap + b_5 * H\_Cap$$

where $P\_Cap = Private\_ind * Capacity$, $H\_Cap = House\_ind * Capacity$.
The regression result is

	Estimate	S.E.	t Value	Pr>|t|
Intercept	35.885	4.111	8.728***	<.001
Capacity	0.659	1.687	0.391	0.695980
Private_ind	20.684	4.672	4.427***	<.001
House_ind	2.293	4.423	0.518	0.604147
P_Cap	7.080	1.947	3.636***	<.001
H_Cap	33.414	1.729	19.323***	<.001

b_4 is the amount to add to b_1 to get the slope for a private room.
b_5 is the amount to add to b_1 to get the slope for a house.