Note 3: In the module we discussed residual sum of squares (RSS) as an error metric for regression, but Turi Create uses root mean squared error (RMSE). These are two common measures of error regression, and RMSE is simply the square root of the mean RSS:

$$RMSE = \sqrt{\frac{RSS}{N}}$$

where N is the number of data points. RMSE can be more intuitive than RSS, since its units are the same as that of the target column in the data, in our case the unit is dollars ($), and doesn't grow with the number of data points, like the RSS does.

(Important note: when answering the question below using Turi Create, when you call the ***linear_regression.create()*** function, make sure you use the parameter ***validation_set=None***, as done in the notebook you download above. When you use regression Turi Create, it sets aside a small random subset of the data to validate some parameters. This process can cause fluctuations in the final RMSE, so we will avoid it to make sure everyone gets the same answer.)

- **What is the difference in RMSE between the model trained with my_features and the one trained with advanced_features?** ***Save this result to answer the quiz at the end.***

✓ Completed

Go to next item

Now, going back to the original dataset, you will build a model using the following features:

```
advanced_features = [
'bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'zipcode',
'condition', # condition of house
'grade', # measure of quality of construction
'waterfront', # waterfront property
'view', # type of view
'sqft_above', # square feet above ground
'sqft_basement', # square feet in basement
'yr_built', # the year built
'yr_renovated', # the year renovated
'lat', 'long', # the lat-long of the parcel
'sqft_living15', # average sq.ft. of 15 nearest neighbors
'sqft_lot15', # average lot size of 15 nearest neighbors
]
```

Note that using copy and paste from this webpage to the Jupyter Notebook sometimes does not work perfectly in some operating systems, especially on Windows. For example, the quotes defining strings may not paste correctly. Please check carefully if you use copy & paste.

- **Compute the RMSE** (root mean squared error) on the test_data for the model using just *my_features*, and for the one using *advanced_features*.

Note 1: both models must be trained on the original sales train dataset, not the one filtered on `sqft_living`.

Note 2: when doing the train-test split, make sure you use seed=0, so you get the same training and test sets, and thus results, as we do.

Note 3: in the module we discussed residual sum of squares (RSS) as an error metric for regression, but Turi Create uses root mean squared error (RMSE). These are two common measures of error regression, and RMSE is simply the square root of the mean RSS:

What you will do

Now you are ready! We are going do three tasks in this assignment. There are 3 results you need to gather along the way to enter into the quiz after this reading.

1. Selection and summary statistics: In the notebook we covered in the module, we discovered which neighborhood (zip code) of Seattle had the highest average house sale price. Now, take the sales data, select only the houses with this zip code, and compute the average price. ***Save this result to answer the quiz at the end.***

2. **Filtering data:** One of the key features we used in our model was the number of square feet of living space ('sqft_living') in the house. For this part, we are going to use the idea of filtering (selecting) data.

- In particular, we are going to use logical filters to select rows of an SFrame. You can find more info in the Logical Filter section of this documentation.
- Using such filters, first select the houses that have 'sqft_living' higher than 2000 sqft but no larger than 4000 sqft.
- What fraction of the all houses have 'sqft_living' in this range? ***Save this result to answer the quiz at the end.***

3. Building a regression model with several more features: In the sample notebook, we built two regression models to predict house prices, one using just 'sqft_living' and the other one using a few more features, we called this set

```
my_features = ['bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', 'floors', 'zipcode
```

Now, going back to the original dataset, you will build a model using the following features:

Congratulations! You passed!

Grade received 100%

Latest Submission Grade 100%

To pass 80% or higher

Retake the assignment in **7h 46m**

Go to next item

1. **Selection and summary statistics:** We found the zip code with the highest average house price. What is the average house price of that zip code?

   1 / 1 point

   Correct

2. **Filtering data:** What fraction of the houses have living space between 2000 sq.ft. and 4000 sq.ft.?

   1 / 1 point

   Correct

3. **Building a regression model with several more features:** What is the difference in RMSE between the model trained with *my_features* and the one trained with *advanced_features*?

   1 / 1 point

   Correct