****

Week 4 Practice with Spark

**Beijing House Pricing Analysis**

ALY 6110 Big Data and Data Management

Student: Yi Yang

Instructor: Daya Rudhramoorthi

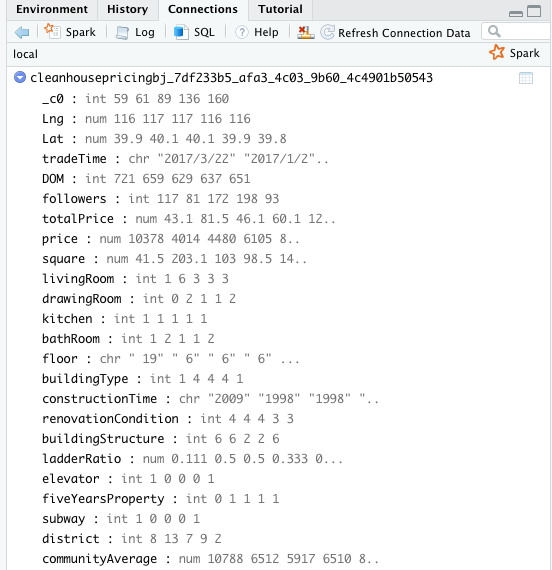
02/17/2021

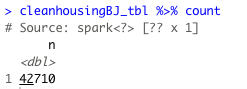
**Introduction**

This assignment is about using Spark to analyze the dataset that is from a Chinese real estate agent company which contains housing price information in Beijing from 2011 to 2017. The purpose for the analysis is to use Sparklyr in RStudio IDE to get to know about this dataset, investigate two business questions that we proposed and generate some insights from it. The questions that whether houses built after 2000 more likely to be sold and how having subway or elevator affect the total price s. By investigating these questions, we can understand more about Beijing’ housing market’s preference and explore what factors that affect the housing price.

**Analysis and Results**

After finishing the data cleansing steps, we obtained the cleaned dataset that contains total 42710 observations and 24 attributes. Then we started installing sparklyr package into RStudio and built the connection with Spark. Next, we imported the data into the system.

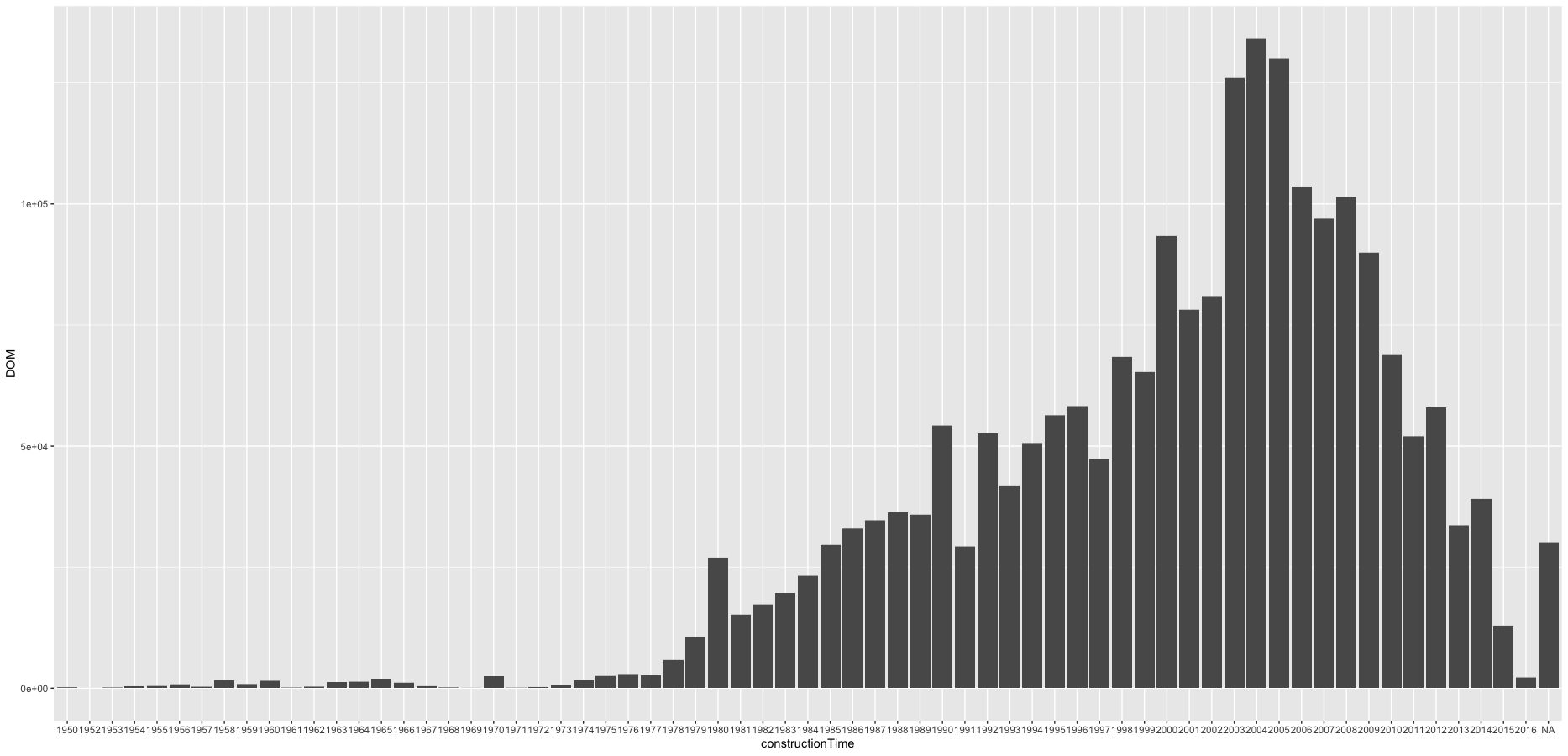




After successfully loading the data file, we started investigating the following business questions:

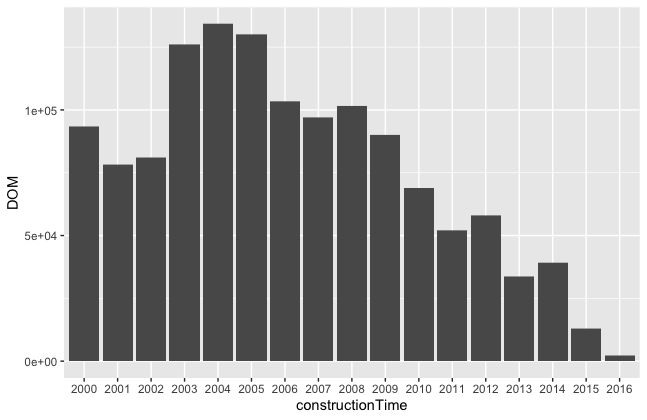
1. Is newer houses (constructionTime >= 2000) more likely to be sold than before 2000?

To explore the answer to this question, we first made a barplot between constructionTime and DOM (days on market) based on all houses. Below is the plot:



*Figure 1.* Barplot of constructionTime vs DOM

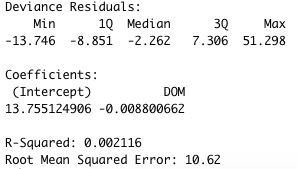
We can observe that houses that are built after 2000 seem to have more days on market than houses that are constructed before 2000. And we can also filter out houses constructed after 2000 using Spark SQL scripts and obtained the plot below.



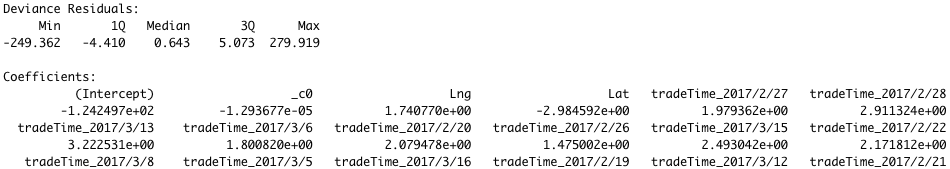
*Figure 2.* Barplot of constructionTime >= 2000 vs DOM

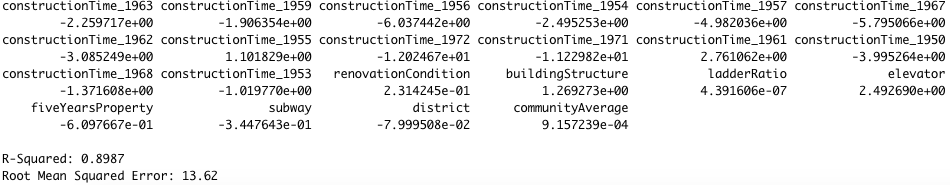
As it shows in Figure 2, we can observe more clearly that houses that are constructed in 2003 to 2005 have the highest number of days on market.

We also built a linear regression model for constructionTime and DOM. After portioning our dataset into train(80%) and validate(20%) sets, we directly used ml\_linear\_regression() algorithm in Spark to run our model.



Moreover, we also built a model for our dependent variable totalPrice and all other variables.

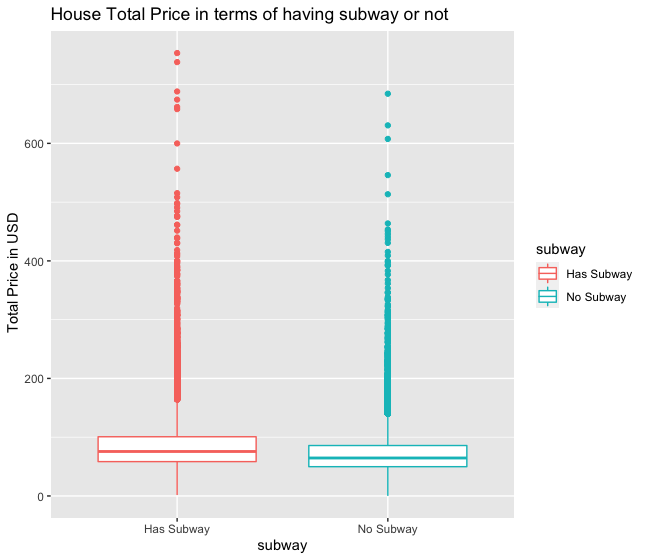




The R-Squared values on both models are relatively small, but the error values for both models are kind of big.

1. Does having no elevator/subway affect houses’ total price?

Furthermore, we are interested at whether having subway will affect the total price by making the plot below.



*Figure 3.* Plot between totalPrice vs subway

Figure 3 tells that there is slightly difference between two options.



*Figure 4.* Plot between totalPrice vs elevator

Compared to Figure 3, the difference of total price in having elevator or not is a bit larger than subway.

**Insights**

Upon finishing all the analysis, we can conclude that houses built in 2003 to 2005 seem to have more days on market. Also, there is a bit difference between the pricing of houses at whether to have elevators or subway, and elevator’s influence is bigger than subway’s. Therefore, findings like these can help real estate companies take these factors into consideration when they make business decisions.