Assignment 2 By Team 4

Salma Bouzid, Shun-Lung Chang, Savitha Singh

1. Explain why it is or why it is not a big data problem.

With respect to the 4Vs (Variety, Velocity, Volume and Veracity) dimentions of big data, we would say it is NOT a big data problem. First of all, the data set, unlike videos or images, is already quite strutured. Of course, the data set still has to be cleaned for further analysis and modeling. But the time spent on data cleaning of this data set could be less than that of other unstrutured data sets. Also, the data set is static rather than streaming data, and hence it could not be a big data problem in terms of velocity. Third, according to the table below, one knows that a data set that has less than 10 gigabytes memory may not be considered as a big data problem. After using `object_size(properties)` in R, one can see that the the memory size of the data set is merely 0.908 GBs, which is far less than 10 GBs. Finally, as can be seen from the introduction page, the "Zestimate" is predicted by millions of machine learning models and the error of these model has been improved consistently. In addition, the other attributes, such as size or location, can gernally be measured by objective observation. As a result, we would say that the data set's veracity has reached a certain extent, and may not need to deal with high inaccurary in some big data problems.

Big	Can't fit in memory on one computer: **>1 TB**
Medium	Fits in memory on a server: **10 GB-1 TB**
Small	Fits in memory on a laptop: **<10 GB**

Big	Many machines, many cores
Medium	Many cores
Small	One core

2. Why is it an analytics problem?

According to the definition of analytics in wikipedia, analytics is to interpret and find valuable pattern in data, and often relies on statistics and computer programming. In this problem, the goal is to find the smallest *log(error)*, which is defined as *log(Zestimate) - log(SalePrice)*. To this end, one has to interpret the data through data viusalization or data transformation, and to build statistical models for predictions. And one often uses programming languages, such as R and Python, to accomplish this task more efficiently and effectively. Therefore, finding the minimal log(error) is a analytics problem.

3. How many data attributes are there?

As can be seen from the result of `colnames(properties)`, the data set contains 58 attributes.

```
colnames(properties)
```

```
 [1] "parcelid"                     "airconditioningtypeid"
 [3] "architecturalstyletypeid"     "basementsqft"
```

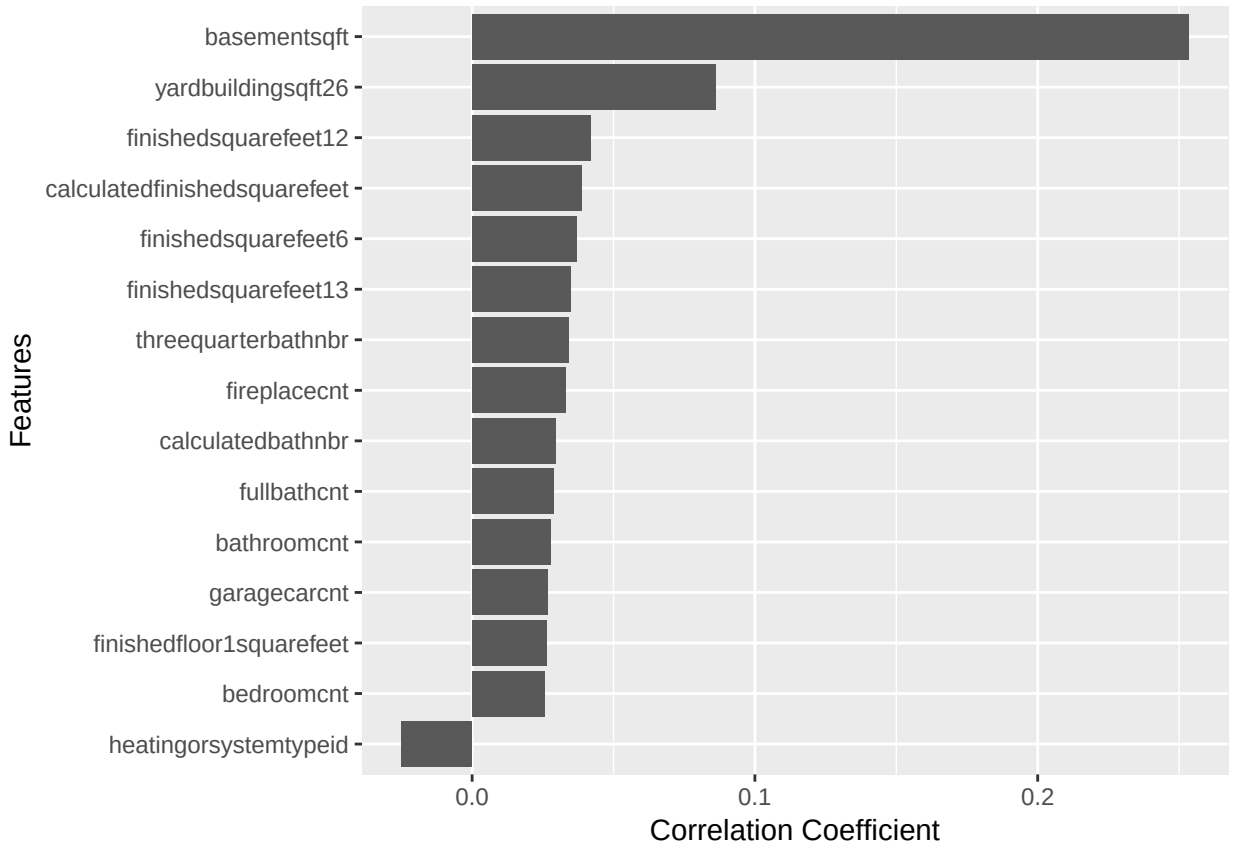
```
 [5] "bathroomcnt"                   "bedroomcnt"
 [7] "buildingclasstypeid"           "buildingqualitytypeid"
 [9] "calculatedbathnbr"             "decktypeid"
[11] "finishedfloor1squarefeet"      "calculatedfinishedsquarefeet"
[13] "finishedsquarefeet12"          "finishedsquarefeet13"
[15] "finishedsquarefeet15"          "finishedsquarefeet50"
[17] "finishedsquarefeet6"           "fips"
[19] "fireplacecnt"                  "fullbathcnt"
[21] "garagecarcnt"                  "garagetotalsqft"
[23] "hashottuborspa"                "heatingorsystemtypeid"
[25] "latitude"                      "longitude"
[27] "lotsizesquarefeet"             "poolcnt"
[29] "poolsizesum"                   "pooltypeid10"
[31] "pooltypeid2"                   "pooltypeid7"
[33] "propertycountylandusecode"     "propertylandusetypeid"
[35] "propertyzoningdesc"            "rawcensustractandblock"
[37] "regionidcity"                  "regionidcounty"
[39] "regionidneighborhood"          "regionidzip"
[41] "roomcnt"                       "storytypeid"
[43] "threequarterbathnbr"           "typeconstructiontypeid"
[45] "unitcnt"                       "yardbuildingsqft17"
[47] "yardbuildingsqft26"            "yearbuilt"
[49] "numberofstories"               "fireplaceflag"
[51] "structuretaxvaluedollarcnt"    "taxvaluedollarcnt"
[53] "assessmentyear"                "landtaxvaluedollarcnt"
[55] "taxamount"                     "taxdelinquencyflag"
[57] "taxdelinquencyyear"            "censustractandblock"
```

4. Identify the type of the 15 attributes you find most relevant in this context

Since the goal of this problem is to find minimal *log(error)*, and we assume that "relevant" means linearly correlated, we show the first 15 attributes that has highest absolute correlation coefficients between logerror below.

First, we picked those features that are numeric, and then used `Hmisc::rcorr()` to get the correlation coefficient matrix. At last we sorted the correlation coefficients between logerror decreasingly, and chose the first 15 items. The result is shown as the barplot.

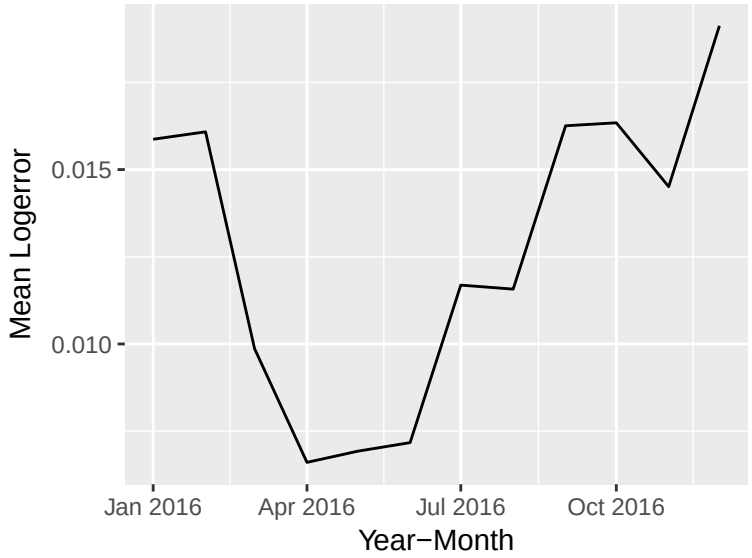
```
numeric_df <- joined_df[, sapply(joined_df, is.numeric), with = FALSE]

corr_mat <- Hmisc::rcorr(as.matrix(numeric_df))

top15 <- abs(corr_mat$r[2, ]) %>%
    sort(decreasing = TRUE) %>%
    .[2:16] %>%
    names()
```


basementsqft
yardbuildingsqft26
finishedsquarefeet12
calculatedfinishedsquarefeet
finishedsquarefeet6
finishedsquarefeet13
threequarterbathnbr
fireplacecnt
calculatedbathnbr
fullbathcnt
bathroomcnt
garagecarcnt
finishedfloor1squarefeet
bedroomcnt
heatingorsystemtypeid
Features
0.0
0.1
0.2
Correlation Coefficient

5. Determine whether the task refers to a supervised or unsupervised learning problem

6. Find out what the standard analysis algorithms are for this analytics problem

7. Provide descriptive summaries of the sample data



8. How are the missings distributed?