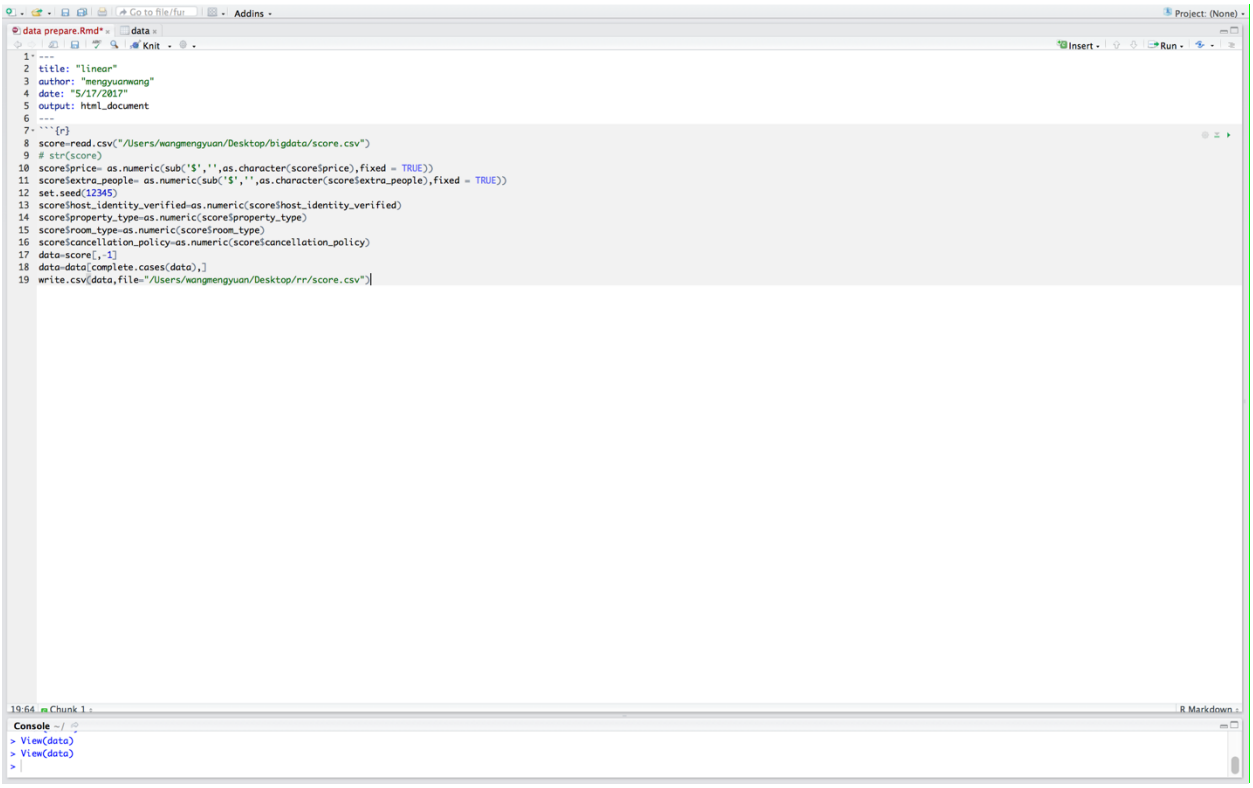
R: data preparation

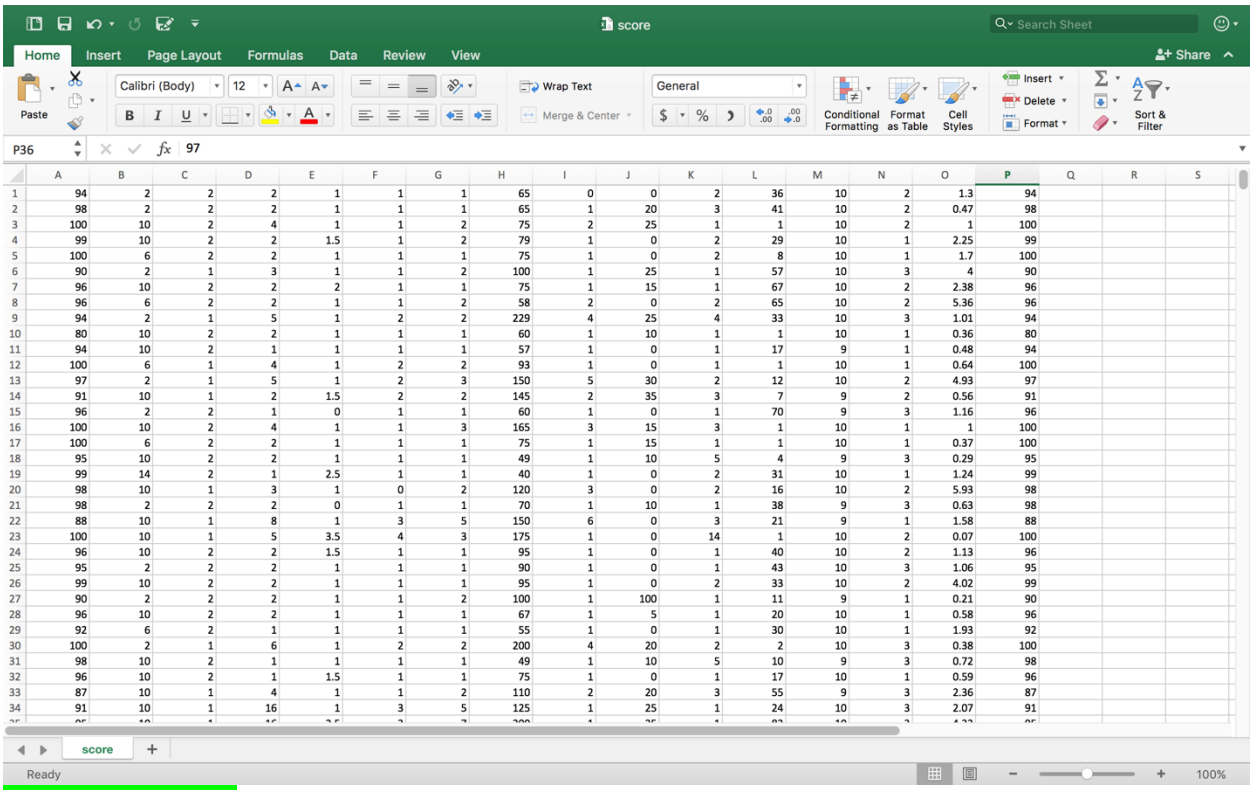
1. Use R to remove the $ and covert 5 factor variables into numeric variables.



```
---
title: "linear"
author: "mengyuanwang"
date: "5/17/2017"
output: html_document
---
```{r}
score=read.csv("/Users/wangmengyuan/Desktop/bigdata/score.csv")
# str(score)
score$price= as.numeric(sub('$','',as.character(score$price),fixed = TRUE))
score$extra_people= as.numeric(sub('$','',as.character(score$extra_people),fixed = TRUE))
set.seed(12345)
score$host_identity_verified=as.numeric(score$host_identity_verified)
score$property_type=as.numeric(score$property_type)
score$room_type=as.numeric(score$room_type)
score$cancellation_policy=as.numeric(score$cancellation_policy)
data=score[,-1]
data=data[complete.cases(data),]
write.csv(data,file="/Users/wangmengyuan/Desktop/rr/score.csv")
```

```
> View(data)
> View(data)
>
```

2. Result save as score.csv



score

Home | Insert | Page Layout | Formulas | Data | Review | View

P36 | 97

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	94	2	2	2	1	1	1	65	0	0	2	36	10	2	1.3	94
2	98	2	2	2	1	1	1	65	1	20	3	41	10	2	0.47	98
3	100	10	2	4	1	1	2	75	2	25	1	1	10	2	1	100
4	99	10	2	2	1.5	1	2	79	1	0	2	29	10	1	2.25	99
5	100	6	2	2	1	1	1	75	1	0	2	8	10	1	1.7	100
6	90	2	1	3	1	1	2	100	1	25	1	57	10	3	4	90
7	96	10	2	2	2	1	1	75	1	15	1	67	10	2	2.38	96
8	96	6	2	2	1	1	2	58	2	0	2	65	10	2	5.36	96
9	94	2	1	5	1	2	2	229	4	25	4	33	10	3	1.01	94
10	80	10	2	2	1	1	1	60	1	10	1	1	10	1	0.36	80
11	94	10	2	1	1	1	1	57	1	0	1	17	9	1	0.48	94
12	100	6	1	4	1	2	2	93	1	0	1	1	10	1	0.64	100
13	97	2	1	5	1	2	3	150	5	30	2	12	10	2	4.93	97
14	91	10	1	2	1.5	2	2	145	2	35	3	7	9	2	0.56	91
15	96	2	2	1	0	1	1	60	1	0	1	70	9	3	1.16	96
16	100	10	2	4	1	1	3	165	3	15	3	1	10	1	1	100
17	100	6	2	2	1	1	1	75	1	15	1	1	10	1	0.37	100
18	95	10	2	2	1	1	1	49	1	10	5	4	9	3	0.29	95
19	99	14	2	1	2.5	1	1	40	1	0	2	31	10	1	1.24	99
20	98	10	1	3	1	0	2	120	3	0	2	16	10	2	5.93	98
21	98	2	2	2	0	1	1	70	1	10	1	38	9	3	0.63	98
22	88	10	1	8	1	3	5	150	6	0	3	21	9	1	1.58	88
23	100	10	1	5	3.5	4	3	175	1	0	14	1	10	2	0.07	100
24	96	10	2	2	1.5	1	1	95	1	0	1	40	10	2	1.13	96
25	95	2	2	2	1	1	1	90	1	0	1	43	10	3	1.06	95
26	99	10	2	2	1	1	1	95	1	0	2	33	10	2	4.02	99
27	90	2	2	2	1	1	2	100	1	100	1	11	9	1	0.21	90
28	96	10	2	2	1	1	1	67	1	5	1	20	10	1	0.58	96
29	92	6	2	1	1	1	1	55	1	0	1	30	10	1	1.93	92
30	100	2	1	6	1	2	2	200	4	20	2	2	10	3	0.38	100
31	98	10	2	1	1	1	1	49	1	10	5	10	9	3	0.72	98
32	96	10	2	1	1.5	1	1	75	1	0	1	17	10	1	0.59	96
33	87	10	1	4	1	1	2	110	2	20	3	55	9	3	2.36	87
34	91	10	1	16	1	3	5	125	1	25	1	24	10	3	2.07	91

score

Ready

Linear Regression
```

1. change the directory
   - `data = sc.textFile('file:/Users/wangmengyuan/Desktop/Bigdata/score.csv')`
   - `data=sqlContext.read.format("libsvm")\`
     `.load("file:/Users/wangmengyuan/Desktop/rr/linear.txt")`
2. #spark-submit --master local[*] --packages com.databricks:spark-csv_2.10:1.2.0 linear.py

```
Last login: Thu May 18 14:47:36 on ttys001
wangmengyuans-MacBook-Pro:~ wangmengyuan$ cd Desktop
wangmengyuans-MacBook-Pro:Desktop wangmengyuan$ cd rr
wangmengyuans-MacBook-Pro:rr wangmengyuan$ spark-submit --master local[*] --packages com.databricks:spark-csv_2.10:1.2.0 linear.py
Ivy Default Cache set to: /Users/wangmengyuan/.ivy2/cache
The jars for the packages stored in: /Users/wangmengyuan/.ivy2/jars
:: loading settings :: url = jar:file:/Users/wangmengyuan/spark/lib/spark-assembly-1.6.1-hadoop1.2.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
com.databricks#spark-csv_2.10 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent;1.0
        confs: [default]
        found com.databricks#spark-csv_2.10;1.2.0 in central
        found org.apache.commons#commons-csv;1.1 in central
        found com.univocity#univocity-parsers;1.5.1 in central
:: resolution report :: resolve 208ms :: artifacts dl 6ms
        :: modules in use:
        com.databricks#spark-csv_2.10;1.2.0 from central in [default]
        com.univocity#univocity-parsers;1.5.1 from central in [default]
        org.apache.commons#commons-csv;1.1 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   3   |   0   |   0   |   0   ||   3   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent
        confs: [default]
        0 artifacts copied, 3 already retrieved (0kB/7ms)
Coefficients:(15,[14],[0.969246768705])
Intercept:2.82736522255
/Users/wangmengyuan/spark/python/lib/pyspark.zip/pyspark/ml/regression.py:123: UserWarning: weights is deprecated. Use coefficients instead.
+-----+--------------------+------------------+
|label|            features|        prediction|
+-----+--------------------+------------------+
| 20.0|(15,[0,1,2,3,4,5,...| 22.212300596652 52|
| 53.0|(15,[0,1,2,3,4,5,...|54.197443963917706|
| 60.0|(15,[0,1,2,3,4,5,...| 60.98217134485275|
| 60.0|(15,[0,1,2,3,4,5,...| 60.98217134485275|
| 60.0|(15,[0,1,2,3,4,5,...| 60.98217134485275|
+-----+--------------------+------------------+
only showing top 5 rows

Model: Root Mean Squared Error = 0.279508321879
wangmengyuans-MacBook-Pro:rr wangmengyuan$
```

Tree Regression

1. change the directory
   - `data = sc.textFile('file:/Users/wangmengyuan/Desktop/rr/score.csv')`
2. #spark-submit --master local[*] --packages com.databricks:spark-csv_2.10:1.2.0 tree.py

```
Last login: Thu May 18 15:14:28 on ttys001
wangmengyuans-MacBook-Pro:~ wangmengyuan$ cd Desktop
wangmengyuans-MacBook-Pro:Desktop wangmengyuan$ cd rr
wangmengyuans-MacBook-Pro:rr wangmengyuan$ spark-submit --master local[*] --packages com.databricks:spark-csv_2.10:1.2.0 tree.py
Ivy Default Cache set to: /Users/wangmengyuan/.ivy2/cache
The jars for the packages stored in: /Users/wangmengyuan/.ivy2/jars
:: loading settings :: url = jar:file:/Users/wangmengyuan/spark/lib/spark-assembly-1.6.1-hadoop1.2.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
com.databricks#spark-csv_2.10 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent;1.0
        confs: [default]
        found com.databricks#spark-csv_2.10;1.2.0 in central
        found org.apache.commons#commons-csv;1.1 in central
        found com.univocity#univocity-parsers;1.5.1 in central
:: resolution report :: resolve 194ms :: artifacts dl 6ms
        :: modules in use:
        com.databricks#spark-csv_2.10;1.2.0 from central in [default]
        com.univocity#univocity-parsers;1.5.1 from central in [default]
        org.apache.commons#commons-csv;1.1 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   3   |   0   |   0   |   0   ||   3   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent
        confs: [default]
        0 artifacts copied, 3 already retrieved (0kB/6ms)
+-----+--------------------+------------------+
|label|            features|       prediction|
+-----+--------------------+------------------+
| 75.0|[2.0,1.0,2.0,1.0,...|78.48314606741573|
| 84.0|[2.0,1.0,2.0,1.0,...|88.28057553956835|
| 87.0|[2.0,1.0,2.0,1.0,...|78.48314606741573|
| 86.0|[2.0,1.0,2.0,1.0,...|88.28057553956835|
| 97.0|[2.0,1.0,2.0,1.0,...|95.99018806214227|
+-----+--------------------+------------------+
only showing top 5 rows

Root Mean Squared Error (RMSE) on test data = 8.39573
wangmengyuans-MacBook-Pro:rr wangmengyuan$
```

Random Forest Regression

1. change the directory
   - `data = sc.textFile('file:/Users/wangmengyuan/Desktop/rr/score.csv')`
2. #spark-submit --master local[*] --packages com.databricks:spark-csv_2.10:1.2.0 randomforest.py

```
Last login: Thu May 18 15:15:44 on ttys001
wangmengyuans-MacBook-Pro:~ wangmengyuan$ cd Desktop
wangmengyuans-MacBook-Pro:Desktop wangmengyuan$ cd rr
wangmengyuans-MacBook-Pro:rr wangmengyuan$ spark-submit --master local[*] --packages com.databricks:spark-csv_2.10:1.2.0 randomforest.py
Ivy Default Cache set to: /Users/wangmengyuan/.ivy2/cache
The jars for the packages stored in: /Users/wangmengyuan/.ivy2/jars
:: loading settings :: url = jar:file:/Users/wangmengyuan/spark/lib/spark-assembly-1.6.1-hadoop1.2.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
com.databricks#spark-csv_2.10 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent;1.0
        confs: [default]
        found com.databricks#spark-csv_2.10;1.2.0 in central
        found org.apache.commons#commons-csv;1.1 in central
        found com.univocity#univocity-parsers;1.5.1 in central
:: resolution report :: resolve 193ms :: artifacts dl 8ms
        :: modules in use:
        com.databricks#spark-csv_2.10;1.2.0 from central in [default]
        com.univocity#univocity-parsers;1.5.1 from central in [default]
        org.apache.commons#commons-csv;1.1 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   3   |   0   |   0   |   0   ||   3   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent
        confs: [default]
        0 artifacts copied, 3 already retrieved (0kB/11ms)
+-----+--------------------+-----------------+
|label|            features|       prediction|
+-----+--------------------+-----------------+
| 93.0|[2.0,1.0,2.0,1.0,...|95.20487723062597|
| 80.0|[2.0,1.0,2.0,1.0,...|78.66629050672455|
| 70.0|[2.0,1.0,2.0,1.0,...|80.93893264921041|
| 80.0|[2.0,1.0,2.0,1.0,...|94.71944955872783|
| 87.0|[2.0,1.0,2.0,1.0,...|94.42225334238051|
+-----+--------------------+-----------------+
only showing top 5 rows

Root Mean Squared Error (RMSE) on test data = 5.59766
wangmengyuans-MacBook-Pro:rr wangmengyuan$
```