Data business analysis

1. What did you do with the data in the context of exploration?

查看数据属性，是否有必要转换数据类型以便于数据分析；查看是否有数据的缺失，重复（检查数据重复函数），异常值的检测，然后读数据进行补全或者删除操作。最后在确定几个关键变量进行描述性，推断性和预测性质的分析。

1. How many entries are in the dataset?

21 variables and 21614 records

1. Was there missing data? Duplications? How clean was the data?
2. Were their outliers or suspicious data?什么是可疑数值

Bedroom outliers

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| id | date | price | bedrooms | bathrooms | sqft\_living | sqft\_lot | floors |
| 1020069017 | 20150327T000000 | 700000.00 | 4 | 1 | 1300 | 1651359 | 1 |
| 722069232 | 20140905T000000 | 998000.00 | 4 | 3.25 | 3770 | 982998 | 2 |
| 2623069031 | 20140521T000000 | 542500.00 | 5 | 3.25 | 3010 | 1074218 | 1.5 |
| 2323089009 | 20150119T000000 | 855000.00 | 4 | 3.5 | 4030 | 1024068 | 2 |
| 3326079016 | 20150504T000000 | 190000.00 | 2 | 1 | 710 | 1164794 | 1 |

1. What did you do to clean the data?
2. What did you find? What intrigued you about the data? Why does that matter?
3. What would your proposed next steps be?

进行全面的相关性分析，以及部分变量的预测分析。制定相应的策略或者提出解决方案。

1. What business questions do you plan to answer with your data mining?

1.最畅销的房屋（大小，房间数目，是否有basemen，景观）是什么类型？或者什么评分段（grade）房子是最受欢迎？什么地区（邮编变量）的房屋是销售最好的？ 什么时候的销售业绩是最好的（data）？

研究这些问题可以帮助房屋销售机构确定销售的目标地区以及户型推荐，了解到最佳销售时机，从而帮助公司制定相应的销售计划和策略，增加利润，以及通过有针对性的房屋推销，包括类型和价格销售，增加客户满意度和忠诚度。

2.研究新建的房子或者最新翻修的房子对于销售有没有相关性？

这样就可以对房屋价值进行评估决定有没有对房子进行翻修后售卖的必要，以获取最大利润

个人作业问题

Explain why and what explicit variables you suggest incorporating. For example, median income by city and state from the census.gov website might be useful for examining home ownership.

是不是我们用到外部数据来对我们研究的数据来进行解释和说明，还是我们本身的数据的去解释，就像我们为什么要选择中位数或着平均值作为我们的分析依据