# Individual Project: Wine Review and Recommendation

Author: Bin Dong

Email: [bindong2@illinois.edu](mailto:bindong2@illinois.edu)

## I. Project Description

This project is aming at two questions

## II. Summary

## A glance at all data

The basic properties of all fields from raw data are:

### Numerics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field | average | var | min | max |
| points | 88.44714 | 9.23996 | 80 | 100 |
| price | 35.36339 | 1682.822 | 4 | 3300 |

Two discoveries from numeric fields: 1. Points(ratings) are ranged from 80 to 100, which are very concentrated on the upper bound of range 0-100. It is better to scale this field 2. There are certain wines with no price.

### Factors

|  |  |
| --- | --- |
| Field | level |
| country | 44 |
| description | 119955 |
| designation | 37980 |
| province | 426 |
| region\_1 | 1230 |
| region\_2 | 18 |
| taster\_name | 20 |
| taster\_twitter\_handle | 16 |
| title | 118840 |
| variety | 708 |
| winery | 16757 |

## Data Processing

1. According to the description of Title column, A title can be discretized as winery+vintage+other information like desigination and province, so I extracted vintage from title according to title constructions. The basic strategy is to replace winery with empty string and parse the first 4-digit string as vintage.