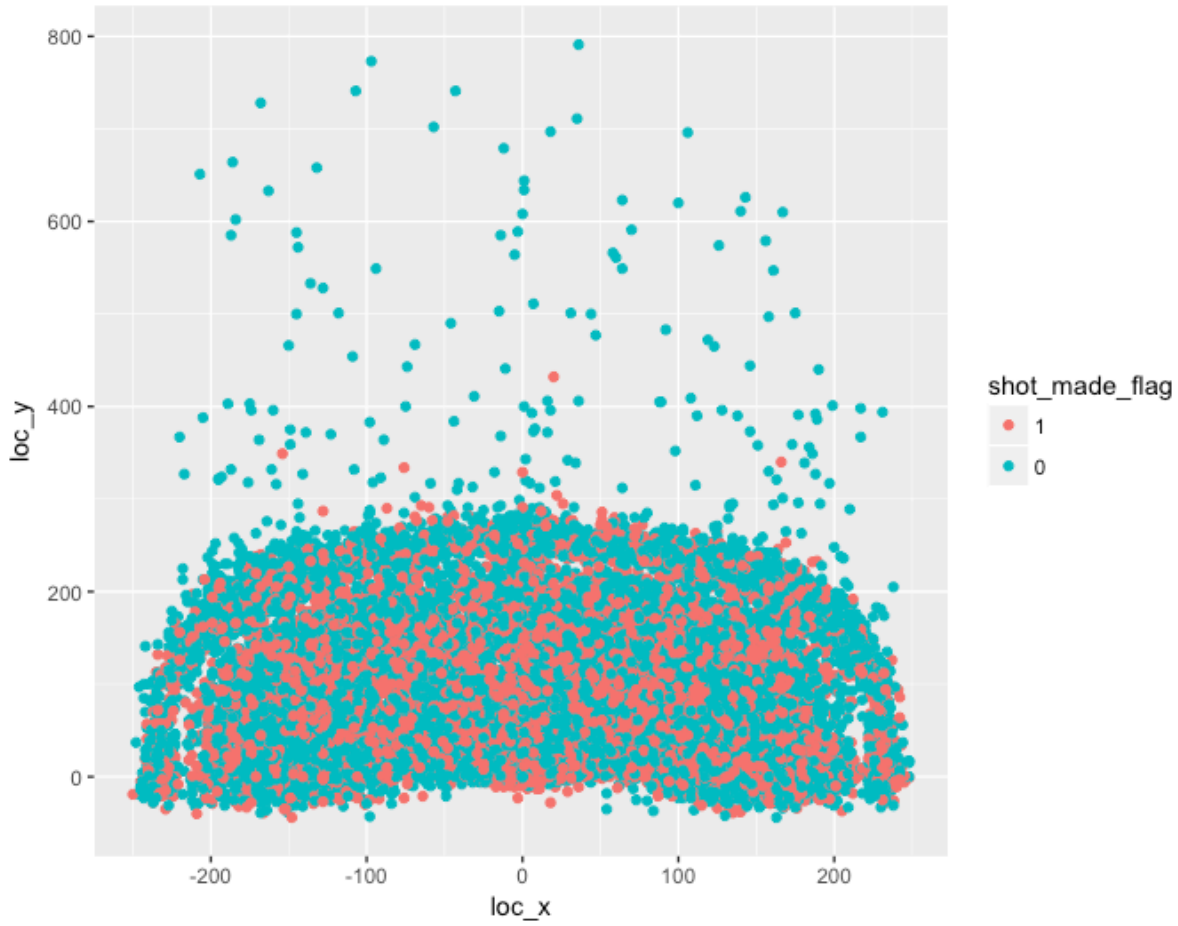
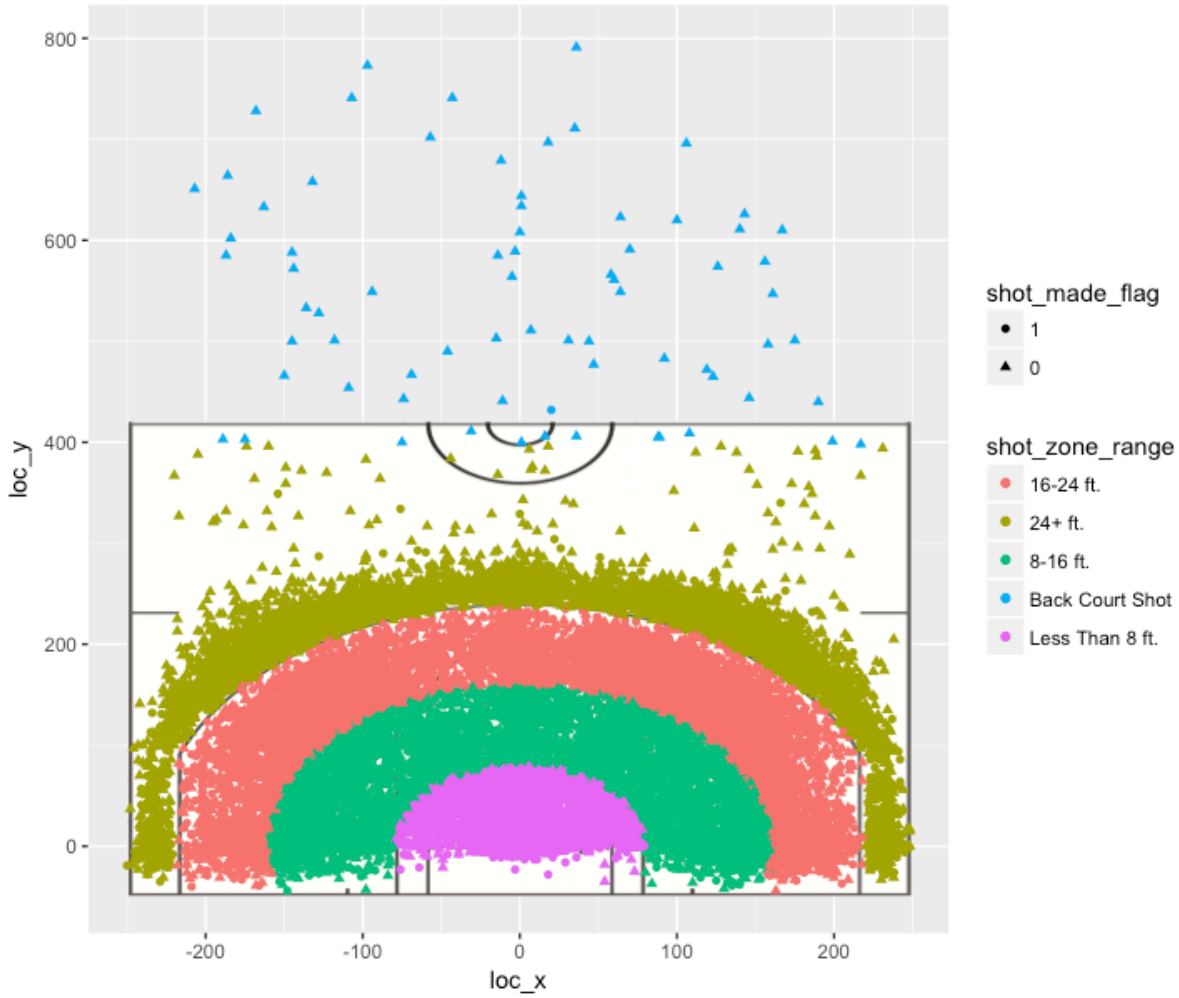
Kobe-Bryant-Shot-Selection

Descriptive Statistics
Total row: 30697

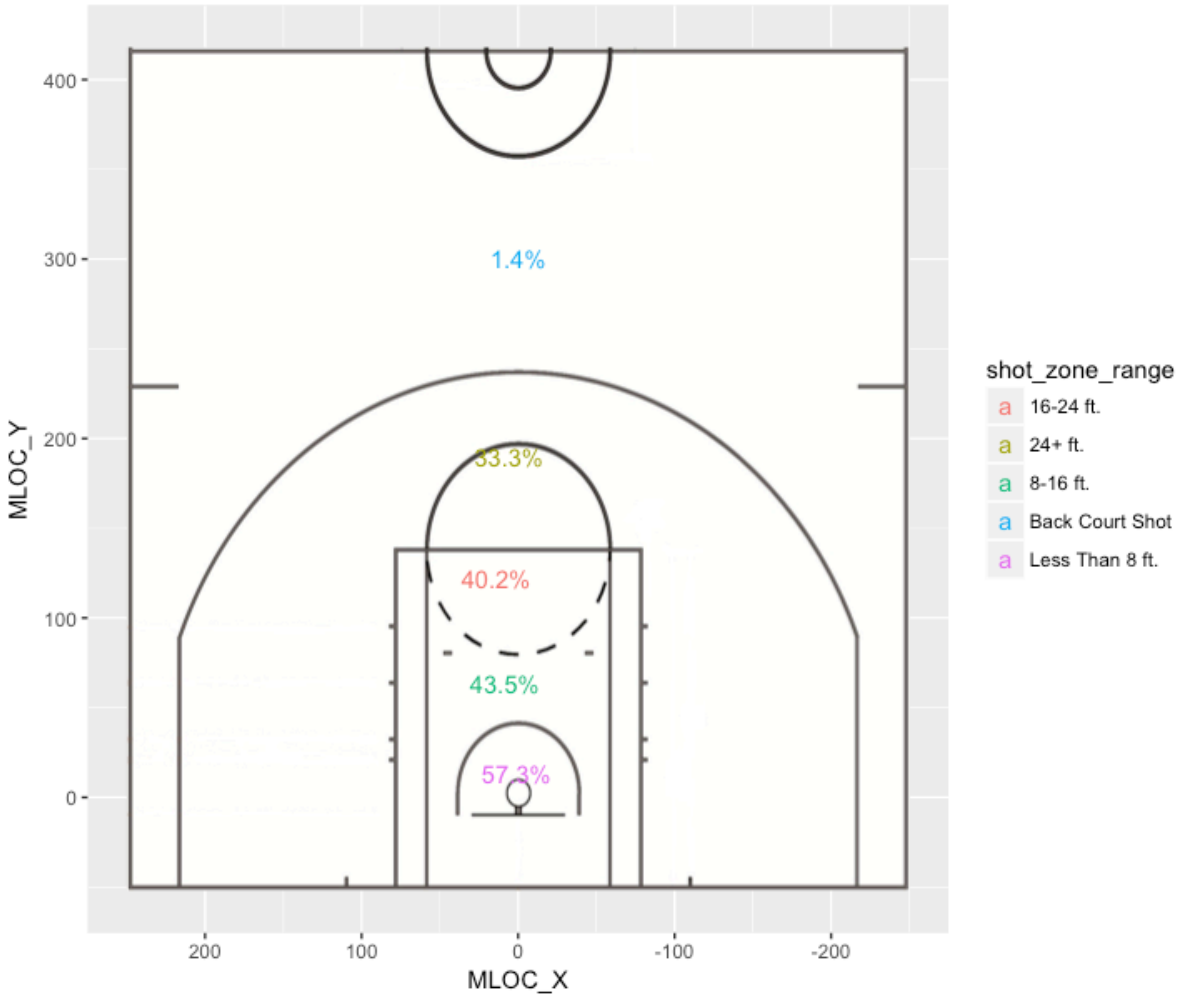
Variables	Unique Count	Remarks
shot_id	30697	Not useful
shot_made_flag	3	Target Variable
game_date	1559	
opponent	33	
game_event_id	620	Useful (Data leakage)
game_id	1559	Useful (Data leakage)
action_type	57	
combined_shot_type	6	
shot_type	2	
shot_zone_area	6	
shot_zone_basic	7	
lat	457	Not useful
lon	489	Not useful
loc_y	457	
loc_x	489	
shot_zone_range	5	
shot_distance	74	
minutes_remaining	12	
seconds_remaining	60	
playoffs	2	Important
season	20	Not useful, same as year
team_name	1	Not useful
team_id	1	Not useful
matchup	74	@, Vs.

1. Attend a simple plot with loc_x and loc_y to observe the geospatial relationship against shot_made_flag.


800
600
400
200
0
loc_y
-200
-100
0
100
200
loc_x
shot_made_flag
1
0



Interpretation: far distance shot beyond 300 is almost all fail to score. Need further segmented with other attributes to find out the relationship.


400
300
200
100
0
MLOC_Y
1.4%
33.3%
40.2%
43.5%
57.3%
200
100
0
-100
-200
MLOC_X
shot_zone_range
16-24 ft.
24+ ft.
8-16 ft.
Back Court Shot
Less Than 8 ft.

A 50% prediction is ranked around 895. This is to benchmark my prediction model's performance.

logistic model only ranking is 966th
xgboost model only ranking is 1006th
0.9 xgboost + 0.1 logistic ranking is 1004th

I submit for the time being. Definitely would further refine later. It is very challenging problem, probably I can't solve this on my own. I would read up and learn how others solve this problem.