Data Analysis : Assignment 2

Ghazal

- Filtered the data for **Paris**
- Selected Hotels and Apartments for accommodation type
- Price is less than 600
- Removed null values from *stars*
- Removed duplicates
- Removed Null values from *rating*, *stars*, and *distance*
- Created log of price, *lnprice*
- Created *highly_rated* if rating >= 4
- Distance is with in 2 miles

```
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
```

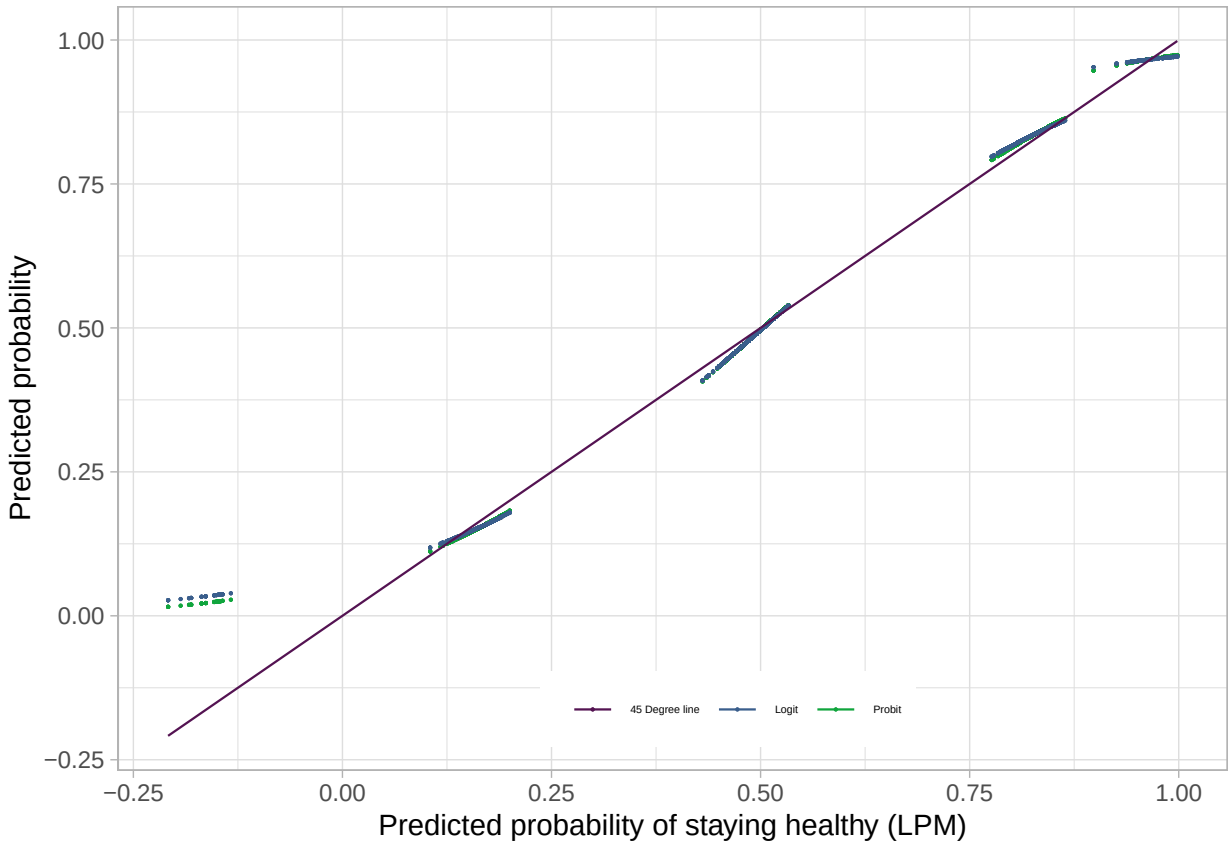
	Mean	SD	Min	Max	Median	P95	N
highly_rated	0.56	0.50	0.00	1.00	1.00	1.00	11397
distance	1.62	0.78	0.10	4.20	1.60	2.90	11397
stars	3.22	0.78	1.00	5.00	3.00	4.00	11397

	Model 1	Model 2	Model 3	Model 4	Model 5
Constant	−0.457**	−4.826**		−2.886**	
	(0.022)	(0.134)		(0.075)	
lspline(stars, c(4))1	0.331**	1.664**	0.312**	0.998**	0.312**
	(0.006)	(0.039)	(0.010)	(0.022)	(0.005)
lspline(stars, c(4))2	0.137**	1.714**	0.275**	0.852**	0.241**
	(0.020)	(0.243)	(0.029)	(0.107)	(0.025)
distance	−0.025**	−0.127**	−0.024**	−0.081**	−0.025**
	(0.005)	(0.028)	(0.005)	(0.017)	(0.005)
Num.Obs.	11 397	11 397	11 397	11 397	11 397

* p < 0.05, ** p < 0.01

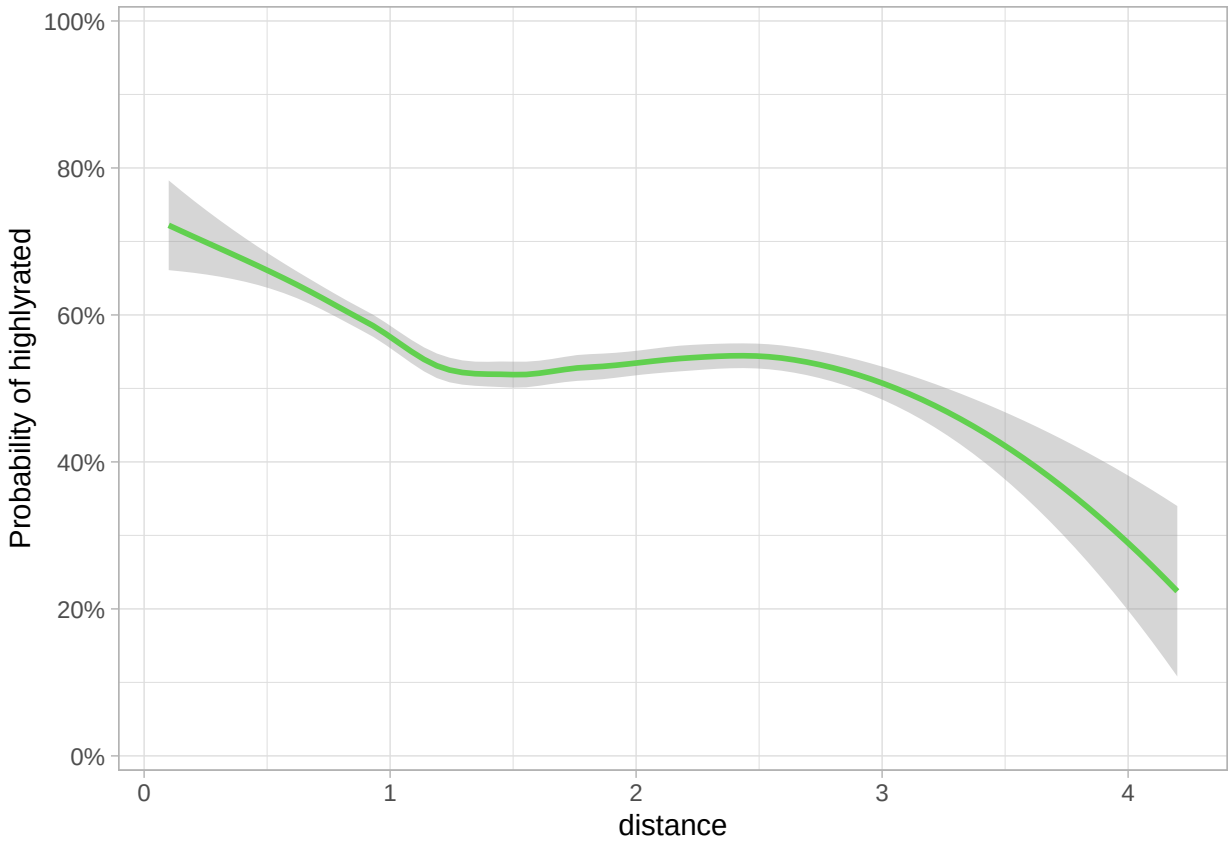
	Model 1	Model 2	Model 3
Constant	−0.457**	−4.826**	−2.886**
	(0.022)	(0.134)	(0.075)
lspline(stars, c(4))1	0.331**	1.664**	0.998**
	(0.006)	(0.039)	(0.022)
lspline(stars, c(4))2	0.137**	1.714**	0.852**
	(0.020)	(0.243)	(0.107)
distance	−0.025**	−0.127**	−0.081**
	(0.005)	(0.028)	(0.017)
Num.Obs.	11 397	11 397	11 397
R2	0.238		
PseudoR2		0.195	0.196

* p < 0.05, ** p < 0.01




```
#distance
g1 <- ggplot(data = data, aes(x=distance, y=highly_rated)) +
  geom_smooth(method="loess", color="3a5e8cFF") +
  scale_y_continuous(expand = c(0.01,0.01),limits = c(0,1), breaks = seq(0,1,0.2), labels = scales::per
  labs(x = "distance",y = "Probability of highlyrated") +
  theme_light()
g1
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
#stars

g1 <- ggplot(data = data, aes(x=stars, y=highly_rated)) +
  geom_smooth(method="loess", color="3a5e8cFF") +
  scale_y_continuous(expand = c(0.01,0.01),limits = c(0,1), breaks = seq(0,1,0.2), labels = scales::per
  labs(x = "stars",y = "Probability of highlyrated") +
  theme_light()
g1
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 3

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 1

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at 3
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius 1

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 1

## Warning: Removed 20 rows containing missing values (geom_smooth).

## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```