# Appendices: New York Restaurants Analyses

Brian Junker

# 9/1/2020

# Contents

Appendix 1. Initial Data Import & Exploration
Appendix 2. Analysis of Apparent Outliers in the EDA Plots
Appendix 3. Regression Analysis – Main Effects Only
Appendix 4. Regression analysis – Two-Way Interactions
Appendix 5. Predicting the price of a restaurant that has very high scores on food, Decor and
Service
Appendix 6. A Table Suitable for Including in a Report

# Appendix 1. Initial Data Import & Exploration

Read the data in, get a general sense of the variables, and make a "pairs" plot (scatterplot matrix) of the numerical variables. Note that "Price" is the response variable.

```
nyc <- read.csv("nyc.csv")</pre>
```

#### str(nyc)

```
## 'data.frame':
                  168 obs. of 7 variables:
  $ Case
##
          : int 12345678910...
                    "Daniella Ristorante" "Tello's Ristorante" "Biricchino" "Bottino" ...
##
  $ Restaurant: chr
              : int 43 32 34 41 54 52 34 34 39 44 ...
##
   $ Price
              : int 22 20 21 20 24 22 22 20 22 21 ...
##
   $ Food
              : int 18 19 13 20 19 22 16 18 19 17 ...
##
  $ Decor
              : int 20 19 18 17 21 21 21 21 22 19 ...
##
   $ Service
   $ East
              : int 000000111...
##
```

summary(nyc)

##	Case		Restaurant			Price		Food	
##	Min.	: 1.00	Length	:168		Min.	:19.0	Min.	:16.0
##	1st Qu.	: 42.75	Class	:charact	cer	1st Qu.	:36.0	1st Qu.	:19.0
##	Median	: 84.50	Mode	:charact	cer	Median	:43.0	Median	:20.5
##	Mean	: 84.50				Mean	:42.7	Mean	:20.6
##	3rd Qu.	:126.25				3rd Qu.	:50.0	3rd Qu.	:22.0
##	Max.	:168.00				Max.	:65.0	Max.	:25.0
##	Dec	or	Serv	ice		East			
##	Min.	: 6.00	Min.	:14.0	Min.	:0.00	0		
##	1st Qu.	:16.00	1st Qu.	:18.0	1st G	u.:0.00	0		
##	Median	:18.00	Median	:20.0	Media	n :1.00	0		
##	Mean	:17.69	Mean	:19.4	Mean	:0.63	1		
##	3rd Qu.	:19.00	3rd Qu.	:21.0	3rd G	u.:1.00	0		
##	Max.	:25.00	Max.	:24.0	Max.	:1.00	0		



We can get a more refined look at the variables with a scatterplot matrix that also includes histograms for each variable. If the histograms revealed especially long tails or wierd outliers, we might want to transform the data, recode or delete outliers, etc.

```
library(psych)
```



(There are lots of other packages, including ggplot, that can produce similar plots. This is really just a convenient illustration.)

The histograms don't suggest any special processing (transformations, etc.) will be needed for the variables, so we can proceed. Note that all the variables seem fairly highly correlated with one another, which makes sense, but also can affect regression results, as we'll learn later in the semester.

One of the main questions for this study is whether restaurants should locate east or west of Fifth Avenue. A pair of boxplots give us a first look at this question:

with(nyc,boxplot(Price ~ East, xlab="East (1 = East of Fifth Avenue)"))



East (1 = East of Fifth Avenue)

# Appendix 2. Analysis of Apparent Outliers in the EDA Plots

To find the two restaurants with modest Service ratings and maximal dinner Prices...

nyc[nyc\$Price==max(nyc\$Price),]

```
##
       Case
                 Restaurant Price Food Decor Service East
## 30
                                      21
                                            20
                                                     20
          30 Harry Cipriani
                                65
                                                            1
## 130
        130
              Rainbow Grill
                                65
                                      19
                                            23
                                                     18
                                                            0
## 132
        132
               San Domenico
                                      23
                                                            0
                                65
                                            22
                                                     22
To find the restaurant with Service = 15...
nyc[nyc$Service==15,]
##
       Case
                      Restaurant Price Food Decor Service East
```

56	56		Nello	54	18	16	15	1
68	68	Zucchero e	e Pomodori	29	17	14	15	1
69	69		Baraonda	37	17	18	15	1
100	100		Ecco-la	25	18	15	15	1
115	115		Lamarca	19	18	9	15	1
	56 68 69 100 115	565668686969100100115115	56 56 68 68 Zucchero 6 69 69 100 100 115 115	56         56         Nello           68         68         Zucchero         e         Pomodori           69         69         Baraonda           100         100         Ecco-la           115         115         Lamarca	56       56       Nello       54         68       68       Zucchero e Pomodori       29         69       69       Baraonda       37         100       100       Ecco-la       25         115       115       Lamarca       19	56         56         Nello         54         18           68         68         Zucchero e Pomodori         29         17           69         69         Baraonda         37         17           100         100         Ecco-la         25         18           115         115         Lamarca         19         18	56       56       Nello       54       18       16         68       68       Zucchero e Pomodori       29       17       14         69       69       Baraonda       37       17       18         100       100       Ecco-la       25       18       15         115       115       Lamarca       19       18       9	56       56       Nello       54       18       16       15         68       68       Zucchero e Pomodori       29       17       14       15         69       69       Baraonda       37       17       18       15         100       100       Ecco-la       25       18       15       15         115       115       Lamarca       19       18       9       15

# Appendix 3. Regression Analysis – Main Effects Only

Here's a very light regression analysis to see how the variables work with one another.

```
summary(lm.0 <- lm(Price ~ . , data=nyc[,-c(1,2)]))</pre>
```

```
## Call:
## lm(formula = Price ~ ., data = nyc[, -c(1, 2)])
##
##
  Residuals:
##
        Min
                  1Q
                        Median
                                     ЗQ
                                              Max
   -14.0465
                        0.0373
                                 3.3942
                                         17.7491
##
             -3.8837
##
## Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
##
                                       -5.102 9.24e-07 ***
##
   (Intercept) -24.023800
                             4.708359
## Food
                 1.538120
                             0.368951
                                        4.169 4.96e-05 ***
## Decor
                 1.910087
                             0.217005
                                        8.802 1.87e-15 ***
                -0.002727
                             0.396232
                                                 0.9945
## Service
                                       -0.007
                 2.068050
                             0.946739
                                                 0.0304 *
## East
                                        2.184
##
   ___
## Signif. codes:
                   0
                     '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.738 on 163 degrees of freedom
## Multiple R-squared: 0.6279, Adjusted R-squared: 0.6187
## F-statistic: 68.76 on 4 and 163 DF, p-value: < 2.2e-16
par(mfrow=c(2,2))
```

```
plot(lm.0)
```



This model, which just has main effects for each of the quantitative predictor variables, suggests some interesting effects on menu prices, and the residual (casewise) diagnostic plots don't show any dramatic misfit,

outliers, influential observations, etc.

From the table of coefficients, it looks like Food and Decor matter a lot for Price, but Service does not. This may be because Service is highly correlated with Food (and with Decor for that matter...).

There is also an effect for being East of Fifth Avenue; this is different from the result we got using only boxplots, because boxplots compare the whole distribution (and so differences have to be true across the distribution of prices) whereas regression analysis basically just looks at means, adjusted for the other variables in the model. Generally when you concentrate inference on the means, you get more dramatic results (because, roughly speaking,  $SE_{mean} = SD_{population}/\sqrt{sample \ size}$ ).

If you are a policy maker (say, you have a lot of money and you are going to open several restaurants), you may care more about the fact that the mean price can be higher East of Fifth Avenue, since on average you can charge a bit more in your restaurants.

On the other hand if you are considering opening just one restaurant, the story of the boxplots may be more important: the price distributions for East vs West restaurants greatly overlap, there's little reason to make a location East of Fifth Avenue a primary concern.

### Appendix 4. Regression analysis – Two-Way Interactions

Just for fun, we'll also try the model that has all main effects and two-way interactions, and we'll compare the two models with likelihood ratio test.

 $summary(lm.1 <- lm(Price ~ .^2, data=nyc[,-c(1,2)]))$ 

```
##
## Call:
## lm(formula = Price ~ .^2, data = nyc[, -c(1, 2)])
##
## Residuals:
##
        Min
                  1Q
                        Median
                                     ЗQ
                                              Max
## -13.7758 -3.5519
                        0.3466
                                 3.3383
                                         17.2584
##
##
  Coefficients:
##
                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)
                  -39.32976
                              42.34684
                                        -0.929
                                                0.35444
## Food
                   2.61252
                               2.34496
                                          1.114
                                                0.26694
## Decor
                   7.26725
                               2.43591
                                          2.983
                                                0.00331 **
## Service
                  -4.68620
                               3.27542
                                        -1.431
                                                0.15450
                                         0.635
## East
                   6.69634
                              10.55070
                                                0.52656
## Food:Decor
                  -0.35758
                               0.13716
                                        -2.607
                                                0.01001 *
## Food:Service
                   0.20733
                               0.15317
                                         1.354
                                                0.17782
## Food:East
                   1.87559
                               0.89562
                                         2.094
                                                0.03785
## Decor:Service
                   0.10665
                               0.09193
                                          1.160
                                                0.24777
## Decor:East
                  -0.34309
                               0.46090
                                        -0.744
                                                0.45775
## Service:East
                  -1.90937
                               0.87262
                                        -2.188
                                                0.03014 *
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.645 on 157 degrees of freedom
## Multiple R-squared: 0.6531, Adjusted R-squared: 0.631
## F-statistic: 29.55 on 10 and 157 DF, p-value: < 2.2e-16
par(mfrow=c(2,2))
```

plot(lm.1)



This model is intersting in that several interactions seem to have coefficients significantly different from zero, and some of the main effects no longer do. The residual plots do not look much better (or worse) than the plots for the main-effects-only model.

As a rule, unless you have a VERY VERY VERY VERY good reason for doing otherwise, when you want to keep an interaction in a model you should also keep the main effects. Thus, if we wanted to keep the Service:East interaction, we should also keep the main variables Service and East, even though neither main effect is significantly different from zero.

However, in this analysis we do not need to worry so much about that, since the likelihood ratio test does not strongly favor the model with interactions; it appears we can "get away" with just the main effects models.

```
anova(lm.0,lm.1,test="LRT")
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Food + Decor + Service + East
##
  Model 2: Price ~ (Food + Decor + Service + East)<sup>2</sup>
##
     Res.Df
                RSS Df Sum of Sq Pr(>Chi)
## 1
        163 5366.5
## 2
        157 5003.4
                     6
                          363.11 0.07694 .
##
   ___
## Signif. codes:
                    0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, we can just stick with the simpler model, lm.0.

Appendix 5. Predicting the price of a restaurant that has very high scores on food, Decor and Service...

## fit lwr upr
## 1 81.40864 69.07858 93.73869

Appendix 6. A Table Suitable for Including in a Report

```
round(summary(lm.0)$coefficients,3)
```

##		Estimate	Std.	Error	t value	Pr(> t )
##	(Intercept)	-24.024		4.708	-5.102	0.000
##	Food	1.538		0.369	4.169	0.000
##	Decor	1.910		0.217	8.802	0.000
##	Service	-0.003		0.396	-0.007	0.995
##	East	2.068		0.947	2.184	0.030