Fried Chicken - Intro to Data Science overview

GWU

2022-09-17

Fried Chicken Pricing

The beginning

At a yelp-highly-rated chicken place one summer, I was waiting for my order, which took **forever** (exaggerated). Without any other better things to do, I re-studied the menu there, trying to understand the pricing structure. See photo:

DataAnalysis_FriedChicken.jpg

I was wondering to myself tons of questions: how can I get the best deal? What is the cost per wing, and cost per drum? Does the restaurant use a formula to determine the pricing?

Bugger! "I am *cursed*. Shouldn't have come here. I might not get out of here alive…" I thought at the time.

You can see the set prices in the photo. We will need to re-format our data, from this *wide* format (often in pivot and summary tables), to a flat, *long* format here. Let us practice our Markdown skill at the same time to tabulate them:

Wings	Drums	Price
5	0	6.69
10	0	12.99
15	0	17.99
20	0	23.99
40	0	45.99
0	3	6.69
0	5	10.99
0	10	19.99
0	15	29.99
3	2	7.99
7	4	16.99
12	6	27.99
20	9	43.99

Let us now enter these as a dataframe in R.

```
# korean-fried-chicken-wing pricing
kfcw <- data.frame(wing=c(-5,10,15,20,40,0,0,0,0,3,7,12,20), drum=c(0,0,0,0,0,3,5,10,15,2,4,6,9), price
```

So there are 13 data points in our dataset. (Notice the use of inline R codes here.)

Exploratory Data Analysis (EDA)

There are a bit of things we typically look at for EDA.

1. Basic statistics
    - mean, s.d., median, range (four spaces at the start of line for proper sub-list indentation)
2. Simple correlations and tests
    - correlation matrix if applicable
    - z-test, t-test, anova test if applicable
    - chi-squared test if applicable
3. Normality
    - QQ-plot
    - boxplot
    - histogram
    - Shapiro-Wilk test
    - …

Basic statistics

```
# str(kfcw)
summary(kfcw)
```

```
  wing            drum            price
```

Min. :-5.0 Min. : 0.00 Min. : 6.7
1st Qu.: 0.0 1st Qu.: 0.00 1st Qu.:11.0
Median : 7.0 Median : 3.00 Median :18.0
Mean : 9.4 Mean : 4.15 Mean :20.9
3rd Qu.:15.0 3rd Qu.: 6.00 3rd Qu.:28.0
Max. :40.0 Max. :15.00 Max. :46.0

When I ran this the first time, the minimum number of wings is **negative 5**. That's obviously a mistake. Have an action plan to fix such non-sensible or missing values. For me, it's just getting some more coffee. After I fixed it, re-run the EDA, and all looks good now.

```
kfcw[1,1]=5  # equal sign "=" and assignment operator "<-" are interchangeable in R
# structure of the data frame kfcw
# str(kfcw)
summary(kfcw)
```

```
  wing            drum            price
```

```
Min. : 0.0 Min. : 0.00 Min. : 6.7
1st Qu.: 0.0 1st Qu.: 0.00 1st Qu.:11.0
Median : 7.0 Median : 3.00 Median :18.0
Mean :10.2 Mean : 4.15 Mean :20.9
3rd Qu.:15.0 3rd Qu.: 6.00 3rd Qu.:28.0
Max. :40.0 Max. :15.00 Max. :46.0
```

Tests (Correlation, ANOVA, . . .)

Since the features/variables are all numerical (quantitative), it makes most sense to check their linear correlations.

```
library("corrplot")
corrmatrix = cor(kfcw)  # more detailed pair-wise correlation test can be obtained from cor.test(kfcw$w
corrplot.mixed(corrmatrix,
               title="Correlation Matrix for Chicken Meal Price",
               mar=c(0,0,1,0) # fixes the position of title
               )
```

RMD-Fried-Chicken-1_files/figure-latex/corr-1.pdf

It is good to see that price has a decent correlation with both wing and drum, with wing seems to have a higher correlation. It is also great that drum and wing do not have too strong a correlation between them. Imagine if each drum I buy, I always will be thrown with 2 dog bones, I will have a hard time knowing if I am really paying for the drum or the dog bones. This is called a problem of collinearity. (We'll check with VIF.)

Normality check/test

It is usually best if all the variables (numerical ones) are normally distributed. This usually does not happen, but we would still like to know overall the distribution is bell-shaped, and not too awkward.

QQ-plot We can first check the QQ-plots for each variable by itself. A straight line means normal distribution.

```
qqnorm(kfcw$price, main = "Price Q-Q Plot", ylab="Price Quantiles ($)")
qqline(kfcw$price)
```

RMD-Fried-Chicken-1_files/figure-latex/normality_qqplot-1.pdf

```
qqnorm(kfcw$wing, main = "Wing Q-Q Plot", ylab="Wing-count Quantiles")
qqline(kfcw$wing)
```

RMD-Fried-Chicken-1_files/figure-latex/normality_qqplot-2.pdf

```
qqnorm(kfcw$drum, main = "Drum Q-Q Plot", ylab="Drum-count Quantiles")
qqline(kfcw$drum)
```

RMD-Fried-Chicken-1_files/figure-latex/normality_qqplot-3.pdf

The data values are not close to normal distribution, but with only 13 data points, that is expected, and it's probably okay.

Boxplot Next we can check the boxplots for a rough visual.

```
# We will learn to use the more powerful ggplot soon, instead of this generic boxplot function
boxplot(kfcw, col=c("red","blue","green"), ylab="count or price($)", main="Boxplots for the three varia
axis(side = 4)
```

RMD-Fried-Chicken-1_files/figure-latex/normality_boxplot-1.pdf

Same conclusion: they do not look like normal, but we'll take it.

Histogram Now histograms:

```
# We will learn to use the more powerful ggplot soon, instead of this generic hist function for histogr
barcolors = c("green", "violet", "orange", "blue", "pink", "red", "yellow", "cyan")
hist(kfcw$price, main = "Histogram for Price distribution", xlab="Price ($)", col=barcolors, breaks = 1
```

RMD-Fried-Chicken-1_files/figure-latex/normality_histogram-1.pdf

```r
hist(kfcw$wing, main = "Histogram for Wing-count", xlab="Wing Count", col=barcolors, breaks = 6)
```

RMD-Fried-Chicken-1_files/figure-latex/normality_histogram-2.pdf

```r
hist(kfcw$drum, main = "Histogram for Drum-count", xlab="Drum Count", col=barcolors, breaks = 6)
```

RMD-Fried-Chicken-1_files/figure-latex/normality_histogram-3.pdf

Shapiro-Wilk test And finally, using shapiro-wilk test:

```r
priceshapiro = shapiro.test(kfcw$price)
wingshapiro = shapiro.test(kfcw$wing)
drumshapiro = shapiro.test(kfcw$drum)
```

The Shapiro-Wilk test p-value on `price` is 0.129.
The Shapiro-Wilk test p-value on `wing` is 0.019.
The Shapiro-Wilk test p-value on `drum` is 0.03.
We'll learn to interpret these soon.

Linear Model

Now if those passed the smell test, we can try build some models. Linear model is first to come to mind. We will model the price with wing and drum as the two independent variables (also called features).

```r
# build a simple linear model (least square fit) of price as a function of everything else.
chicklm = lm(price ~ ., data=kfcw)
summary(chicklm)
```

Call: lm(formula = price ~ ., data = kfcw)

Residuals: Min 1Q Median 3Q Max -1.3066 -0.3005 -0.0471 0.1524 1.8448

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.7775 0.5038 1.54 0.15
wing 1.1630 0.0255 45.67 6.1e-13 ***drum 2.0120 0.0620 32.44 1.8e-11*** — Signif. codes: 0 ‘***’ 0.001 ’***’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 0.939 on 10 degrees of freedom Multiple R-squared: 0.996, Adjusted R-squared: 0.995 F-statistic: 1.16e+03 on 2 and 10 DF, p-value: 1.43e-12

As we can see, the R^2 (using LaTeX formatting here) value of the linear model is 0.9957, which shows the prices are set with rather strict per-piece-structure.

If we use the (default 95%) confidence intervals of the coefficients as shown here:

```
coeffconfint = confint.lm(chicklm)
coeffconfint
```

```
         2.5 % 97.5 %
```

(Intercept) -0.345 1.90 wing 1.106 1.22 drum 1.874 2.15

Findings

We find that:

1. Each piece of wing cost about $1.11 to $1.22
2. Each piece of drum stick cost about $1.87 to $2.15
3. The intercept of about $-0.345 to $1.9, or average of $0.78 probably represent the base per-order or box/bag charge.

Check:
We should always check for multi-collinearity in linear models. Here are the **Variance Inflation factors VIF**.

```
library("faraway") # faraway library is one of them has a vif function
# VIF check the collinearity issues between different variables/features in a (linear) model
vif(chicklm)
```

wing drum 1.19 1.19 With VIF values less than 5, we can safely conclude there is not much collinearity concerns in the dataset.

Plot 3D

Visual is always king in data science. Let's get a bit fancy.

```
library("plot3D")
# reference from http://www.sthda.com/english/wiki/impressive-package-for-3d-and-4d-graph-r-software-an
# x, y, z variables
x <- kfcw$wing
y <- kfcw$drum
z <- kfcw$price
# Compute the linear regression (z = ax + by + d)
# chicklm <- lm(z ~ x + y)
# predict values on regular xy grid
grid.lines = 20
x.pred <- seq(min(x), max(x), length.out = grid.lines)
y.pred <- seq(min(y), max(y), length.out = grid.lines)
xy <- expand.grid( wing = x.pred, drum =y.pred)
z.pred <- matrix(predict.lm(chicklm, newdata = xy), nrow = grid.lines, ncol = grid.lines)
# fitted points for droplines to surface
fitpoints <- predict.lm(chicklm)
# scatter plot with regression plane
scatter3D(x, y, z, pch = 18, cex = 2, theta = 5, phi = 20, ticktype = "detailed", xlab = "wing", ylab =
```

RMD-Fried-Chicken-1_files/figure-latex/3dscatter2-1.pdf

Conclusion

So what is the best deal???

First bite

Let me compare the predicted values from the model with the actual values. In other words, look at the residual values.

```
kfcw$fitval = chicklm$fitted.values
kfcw$residuals = chicklm$residuals
kfcw
```

wing drum price fitval residuals 1 5 0 6.69 6.59 0.0976 2 10 0 12.99 12.41 0.5827 3 15 0 17.99 18.22 -0.2322 4 20 0 23.99 24.04 -0.0471 5 40 0 45.99 47.30 -1.3066 6 0 3 6.69 6.81 -0.1236 7 0 5 10.99 10.84 0.1524 8 0 10 19.99 20.90 -0.9077 9 0 15 29.99 30.96 -0.9678 10 3 2 7.99 8.29 -0.3005 11 7 4 16.99 16.97 0.0235 12 12 6 27.99 26.81 1.1846 13 20 9 43.99 42.15 1.8448

Ah, the residual is most negative with the 40-wing combo at -$1.31. That saves me a can of soda. I should go for that and kill myself.

But if I am wiser, and insist on having a one-person portion:

1. 5-piece wing is at +$0.10 (paying a dime too much!) - Not good.
2. 3-piece drum is at -$0.12, saving a dozen pennies, sounds good.
3. The 3-wing-2-drum deal (at -$0.30) is the best value for me!

Double dip

Hold on! If we plan to do this regularly, shouldn't we inspect the percentage savings instead?

```
kfcw$res_2_p = chicklm$residuals/kfcw$price*100
kfcw
```

wing drum price fitval residuals res_2_p 1 5 0 6.69 6.59 0.0976 1.458 2 10 0 12.99 12.41 0.5827 4.486 3 15 0 17.99 18.22 -0.2322 -1.291 4 20 0 23.99 24.04 -0.0471 -0.196 5 40 0 45.99 47.30 -1.3066 -2.841 6 0 3 6.69 6.81 -0.1236 -1.848 7 0 5 10.99 10.84 0.1524 1.386 8 0 10 19.99 20.90 -0.9077 -4.541 9 0 15 29.99 30.96 -0.9678 -3.227 10 3 2 7.99 8.29 -0.3005 -3.761 11 7 4 16.99 16.97 0.0235 0.139 12 12 6 27.99 26.81 1.1846 4.232 13 20 9 43.99 42.15 1.8448 4.194

The new metric here tells me:

1. The 10-drum deal at -4.5% is best percentage-wise.
2. The 3-wing-2-drum deal at -3.76% is next.

To me, putting all these different metrics and criteria together, the 3-wing-2-drum is my winner.

Case closed. Mic dropped.

Confession

Oh wait… I forgot that I am a pescatarian (vegetarian and fish) for over a decade! I just have these cravings for wings once in a while. What if I add a sin-ful parameter as a penalty term, how much should I order next time to minimize my sin (Root-Mean-Squared-Sinfulness).

To be continued…

But yes, the wings there were really really good. Totally worth the wait.

Reference

APA Style preferred

1. Yelp
2. Google
3. My wallet