# BZAN 535 - Homework 4

*Kevin Garder*

*September 20, 2016*

```
## ------------------------------------------------------------------------
```

```
## data.table + dplyr code now lives in dtplyr.
## Please library(dtplyr)!
```

```
## ------------------------------------------------------------------------
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: rJava
```

```
## Loading required package: xlsxjars
```

# Question 1

We wish to measure the impact of having an item appear in the Kroger weekly advertisement. You will study five 26 oz. Ragu pasta sauce with upc in (3620000250, 3620000300, 3620000350, 3620000441, 3620000446). For three Kentucky stores (333, 352, 377), these five upcs were all displayed in the weekly flyer (Interior Page Feature) for 18 weeks, with 14 of those weeks in year 2, i.e., with weeks > 52. This information is contained in the table carbo_causal_lookup.

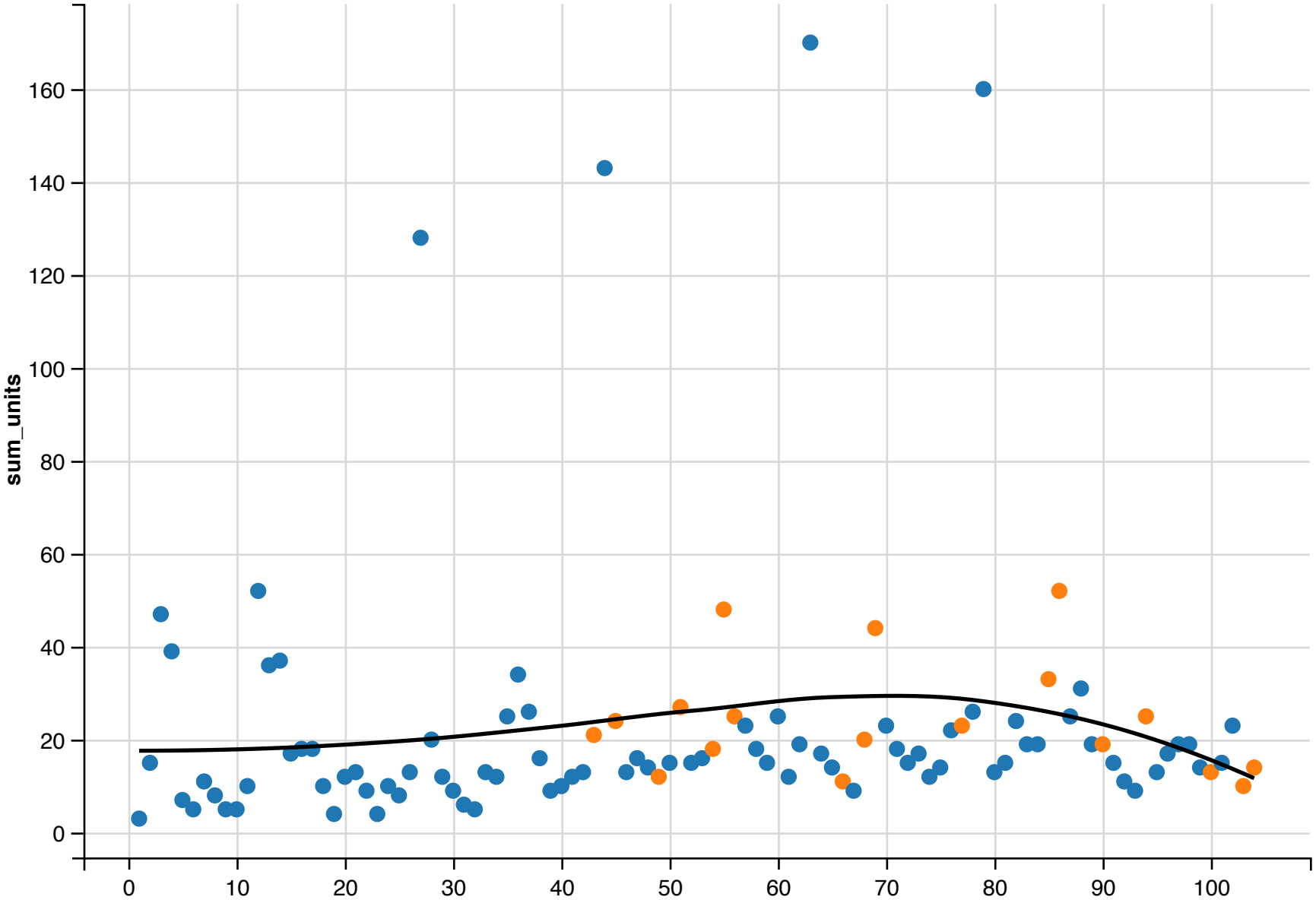
a. Show a query that will capture all the transaction data for these upcs and stores,aggregating the data by week, and computing the total unit sales, summed across stores and upcs. (You may find it useful to include other columns later, but these are not required for answering 1a.)

```
select week, sum(units) as sum_units, sum(dollar_sales)/sum(units) as avg_price
from carbo_transactions ct
where ct.upc in (3620000250,3620000300,3620000350,3620000441,3620000446)
and ct.store in (333,352,377)
group by week;
```

b. Using the carbo_causal_lookup table, identify the 18 weeks where these upcs appear in the interior of the weekly flyer.

```
select *
from carbo_causal_lookup
where upc in (3620000250,3620000300,3620000350,3620000441,3620000446)
and store in (333,352,377)
and feature_desc in ('Interior Page Feature','Interior Page Line Item')
group by week;
```

c. Create a plot of total unit sales by week, highlighting the 18 weeks identified in b. Annotate any other weeks as seems relevant, based on other information available from the transactions or causal_lookup tables.




d. Discuss what you can learn from the plot. Using just the plot, does there appear to be any indication that the weekly advertisements help sales of these products?

The plot does not seem to indicate that the weekly advertizements help sales. The 18 weeks identified in b (colored light blue) do not seem to have higher sales than the other 86 weeks.

# Question 2

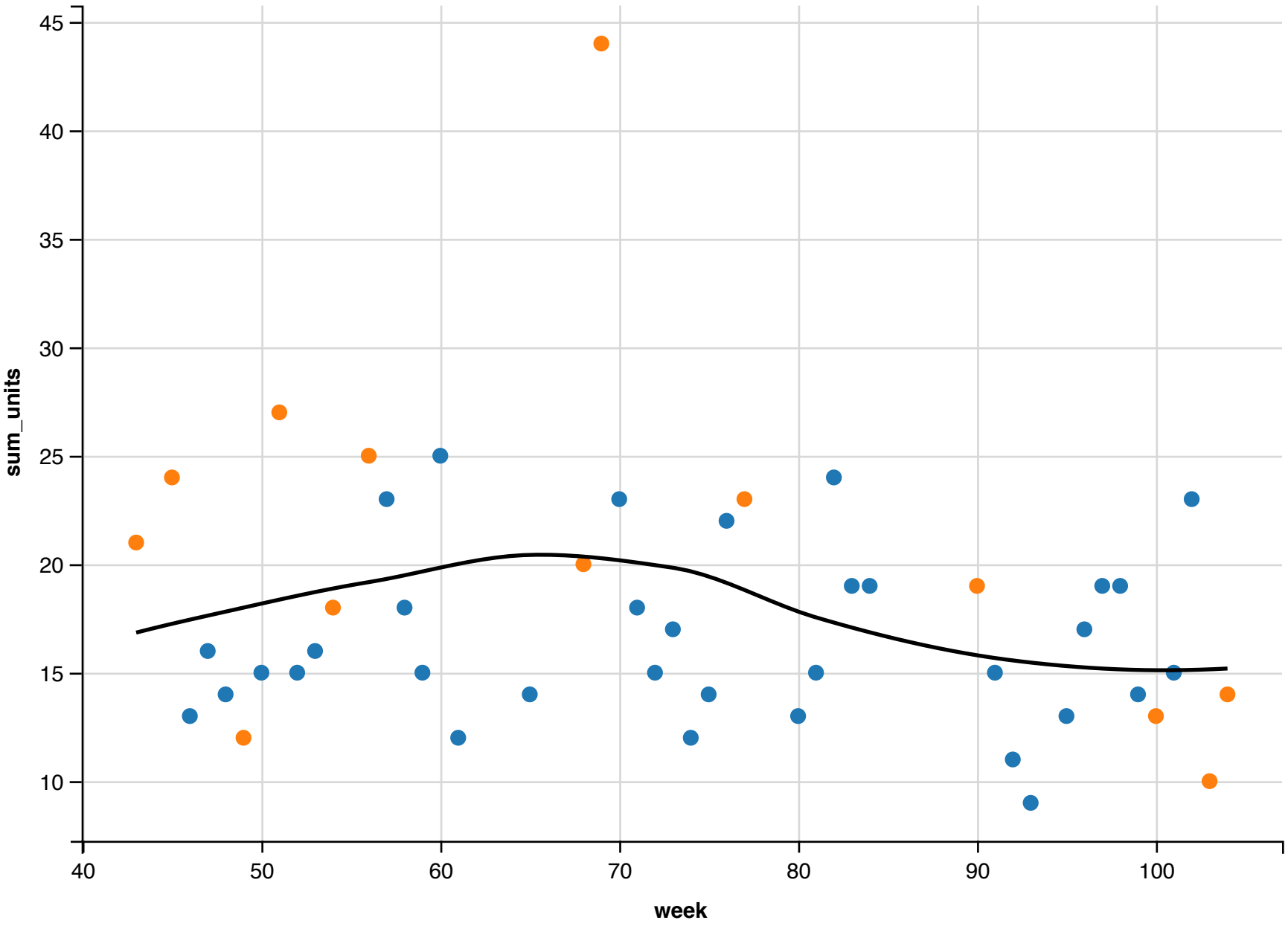
Propose a paired sample method to evaluate the effect of the flyer.

a. First, identify the criteria you would propose for choosing pairs of weeks. What should they have in common?

From the plot we see there is correlation between week and unit sales, so in subsetting our sample we want to look at weeks that are similar to each other so we can detect differences that are due to the flyer. Products are advertized in the flyer intermittently from week 47-104, so to pair the data we focus on unit sales within this period. This gives us subset with 18 observation with advertizement and 44 observations without advertizement.

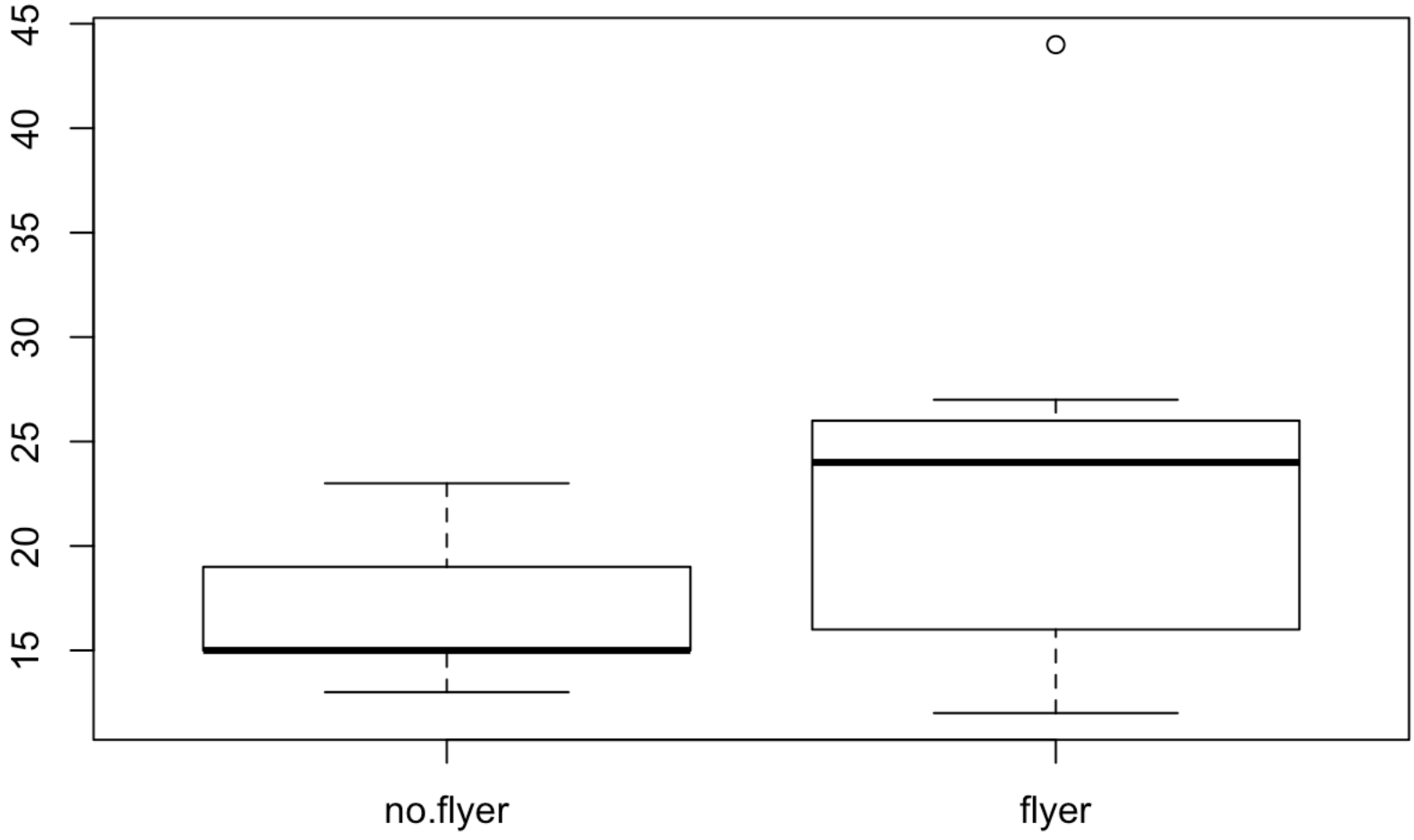
We want to pair observations by criteria that effect unit sales. Since all `5` products are Ragu pasta sauce and all `3` stores are in Kentucky, we will assume UPC and store are not significanly influencing sales. We expect a strong relationship between average price and unit sales, and we also suspect that seasonaly could influences unit sales.

Below is a scatterplot of the subset of weeks we'll choose pairs from.




b. Second, select weeks to be paired with at least 10 of the 18 weeks from 1b. Explain the rationale for your choices, arguing why these matches were provided and why any weeks not used were omitted.

Most of the "with advertisement" observations are lagged one week behind one of the "without advertizement" observations, so we increment the week by 1 for the "with advertizement" observations and are able to obtain 12 pairs. Below is pair of boxplots for the resulting sample.




c. Conduct a paired sample hypothesis test. Write out the null and alternative hypotheses very clearly in the language of the problem. (Make sure that both hypotheses include the word ‘average’ or ‘expected’.) Show all the details of the test using computer output. Interpret the p-value clearly and state your conclusion.

We test the hypotheses:
$H_0$: average unit sales without flyer - average of unit sales with flyer = 0
$H_1$: average of unit sales without flyer - average of unit sales with flyer < 0

```
t.test(pairs$no.flyer, pairs$flyer, alternative = "less", paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  pairs$no.flyer and pairs$flyer
## t = -1.9646, df = 6, p-value = 0.04854
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.07004045
## sample estimates:
## mean of the differences
##               -6.428571
```

d. Comment about the correlation between the pairs of your data, and what this indicates about the effectiveness of your pairing.

```
cor(pairs$no.flyer, pairs$flyer)
```

```
## [1] 0.653367
```

The correlation $\rho = 0.653367$ is somewhat strong and suggests that the pairing strategy was reasonably effective.

# Quesion 3

One might also try to use more of the data and give no thought to pairing, just treating the data as two independent samples.

a. Out of the 104 -18 weeks you did not have the products displayed in the inner pages, which should be included for this comparison. Justify your choices of exclusions and inclusions.

Here we simply choose a random sample from weeks without advertisements in order to obtain 18 pairs.

b. Conduct an independent sample test using the Interior Page Feature weeks and the other weeks you identified in 3a. Show the test results using output from JMP, interpret the p-value, and state your conclusion.

```
t.test(control$sum_units, treatment$sum_units, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  control$sum_units and treatment$sum_units
## t = -1.9036, df = 32.998, p-value = 0.03286
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.8137589
## sample estimates:
## mean of x mean of y
##  17.05556  24.38889
```

Using independent samples we get a p-value of $0.03286$ so we are able to reject the null hypothesis at the 95% confidence level.