Ben & Jerry: Ice-Cool Stats

(Student: Vinh Luong - 442069)

1. COOL TOPIC

With the retreat of the Chicago winter, it is now pleasant enough to look at a kind of data set that would have made us shiver just a few weeks ago: 21,974 purchases of Ben & Jerry's ice-creams throughout the country.

The key question we will be tackling today is: **what makes people buy more expensive ice-creams**? Why? More expensive ice-creams create *dolce vitae* for everyone around, from the buyers who can afford them to Messrs. Ben Cohen and Jerry Greenfield and their spouses and heirs and heiresses and their shareholders and the causes and politicians they support.

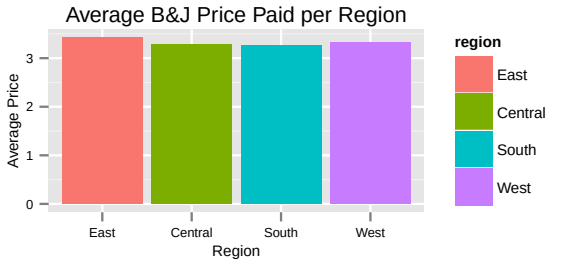
But before we can take a bite at any cool, sweet data, we must do some serious pre-processing work, which tastes rather like a disappointingly sour lemon...

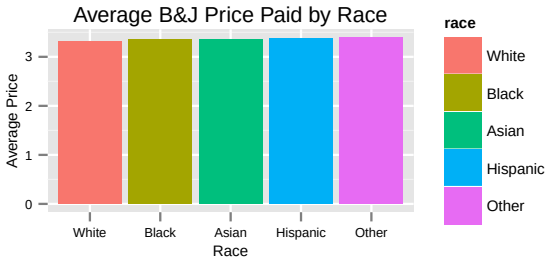
```
# Zoom into some variables of interest and do some basic data-cleaning
# and feature-creation
# (skipping: male_head_occupation, female_head_occupation, household_composition;
# because these are reflected by other variables)
# (skipping: total_spent; don't know what it means)
benjer <- benjer[, .(quantity, price_paid_deal, price_paid_non_deal,
                     size1_descr, flavor_descr, formula_descr,
                   coupon_value, promotion_type,
                   household_size, household_income, age_and_presence_of_children,
                   male_head_employment, female_head_employment, male_head_education,
                   female_head_education, marital_status, race,
                   hispanic_origin, region, type_of_residence, kitchen_appliances,
                   tv_items, household_internet_connection)]

# Transform Household Income ordinal values to numeric values (intervals' midpoints)
household_income_ordinal = c(3, 4, 6, 8, 10, 11, 13, 15, 16, 17, 18, 19, 21, 23, 26,
                             27, 28, 29, 30)
household_income_numeric = c(mean(0, 5000), mean(5000-7999), mean(8000, 9999),
                             mean(10000, 11999), mean(12000, 14999), mean(15000, 19999),
                             mean(20000, 24999), mean(25000, 29999), mean(30000, 34999),
                             mean(35000, 39999), mean(40000, 44999), mean(45000, 49999),
                             mean(50000, 59999), mean(60000, 69999), mean(70000, 99999),
                             mean(100000, 124999), mean(125000, 149999),
                             mean(150000, 199999), 200000)

# Fill NA values
benjer$promotion_type[is.na(benjer$promotion_type)] <- 0
benjer$tv_items[is.na(benjer$tv_items)] <- 0

# Put Hispanic Origin into Race
benjer$race[(benjer$race == 4) & (benjer$hispanic_origin == 2)] <- 5

# Create New Input Features
```

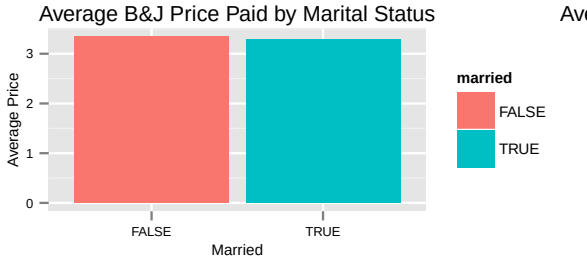
```
benjer[, `:=`(price_paid_per_1 = (price_paid_deal + price_paid_non_deal) / quantity,
              coupon_used = (coupon_value > 0),
              coupon_value_per_1 = coupon_value / quantity,
              size1_descr = factor(size1_descr),
              flavor_descr = factor(flavor_descr),
              formula_descr = factor(formula_descr),
              promotion_type = factor(promotion_type, levels = 0 : 4,
                                      labels = c("NONE", "StoreFeature", "StoreCoupon",
                                                 "ManufacturerCoupon", "OtherDeal")),
              household_income = mapvalues(household_income,
                                           household_income_ordinal,
                                           household_income_numeric),
              children_under_18 = age_and_presence_of_children < 9,
              male_head_employed_full_time = (male_head_employment == 3),
              female_head_employed_full_time = (female_head_employment == 3),
              male_head_graduated_college = (male_head_education >= 5),
              female_head_graduated_college = (female_head_education >= 5),
              married = (marital_status == 1),
              race = factor(race, levels = 1 : 5,
                            labels = c("White", "Black", "Asian", "Hispanic", "Other")),
              region = factor(region, levels = 1 : 4,
                              labels = c("East", "Central", "South", "West")),
              one_family_house = (type_of_residence == 1),
              microwave = kitchen_appliances %in% c(1, 4, 5, 7),
              dishwasher = kitchen_appliances %in% c(2, 4, 6, 7),
              garbage_disposal = kitchen_appliances %in% c(3, 5, 6, 7),
              cable_tv = (tv_items > 1),
              internet = (household_internet_connection == 1))]

# Remove Old Features
benjer[, c("quantity", "price_paid_deal", "price_paid_non_deal", "coupon_value",
           "household_size", "age_and_presence_of_children", "male_head_employment",
           "female_head_employment", "male_head_education", "female_head_education",
           "marital_status", "hispanic_origin", "type_of_residence", "kitchen_appliances",
           "tv_items", "household_internet_connection") := NULL]
```

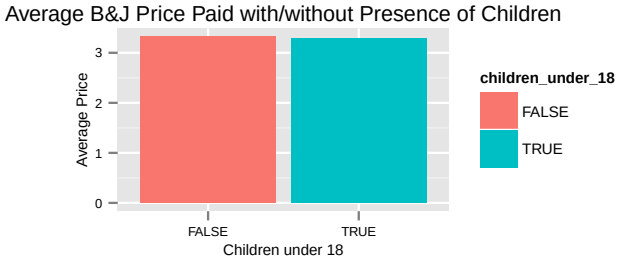
Phew... over 48 hours have elapsed since I started cleaning the data. Along the way, I managed to annoy a few folks on StackOverflow and report a *data.table* bug.

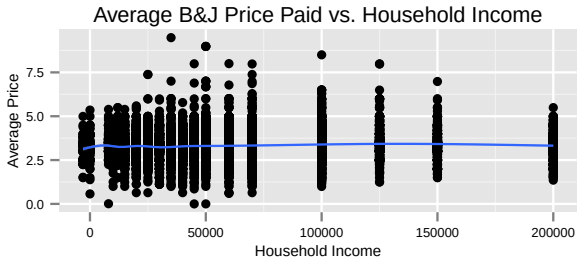
Now we are ready to look at some colorful plots of the average price paid for one ice-cream against various factors:





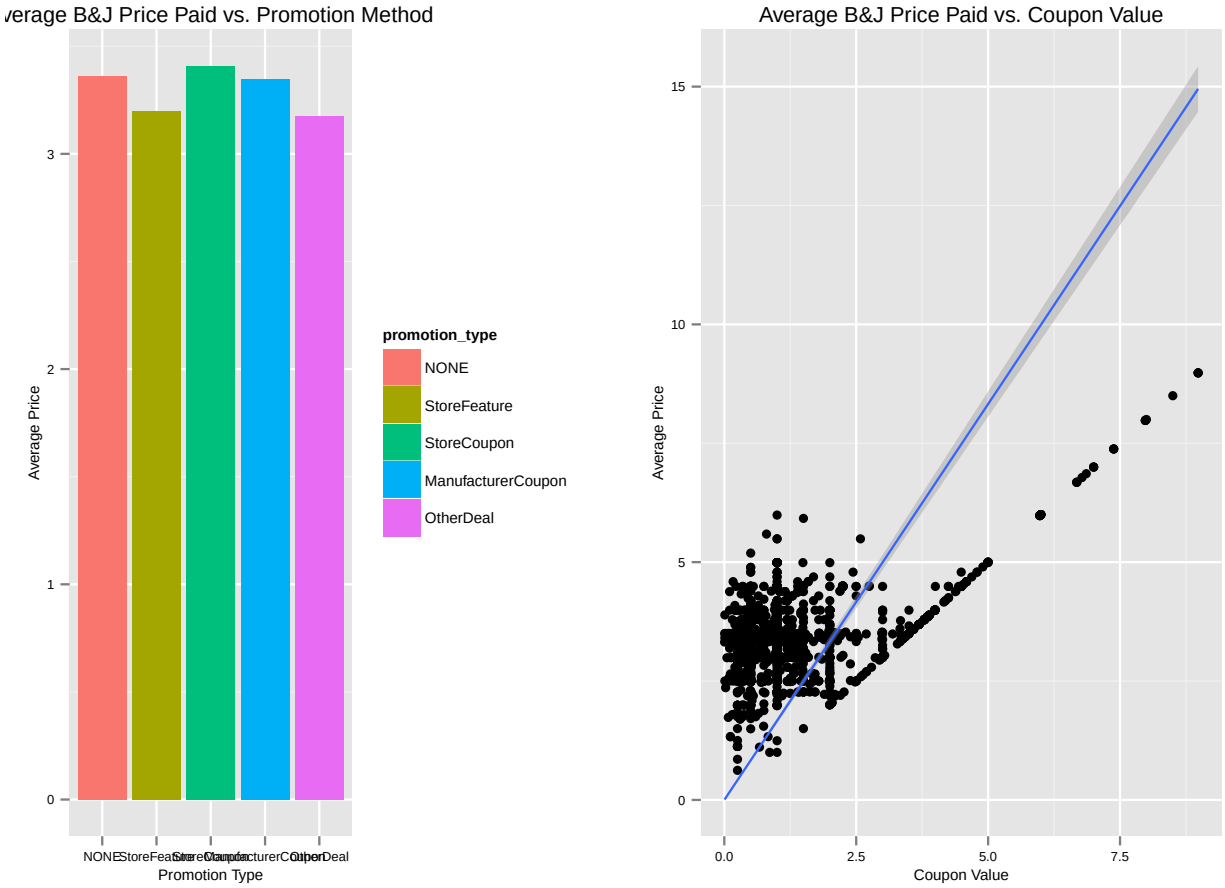






From a first look at the above 5 plots there seems to be little difference among the various ways to split the data. It seems, rather uninterestingly, *when people have decided that they want ice-creams at Ben & Jerry's*, they are willing to pay pretty much the same amount on average, regardless of their U.S. home region, their races, whether they are married, whether they have kids, or how much money they make. Americans seem to appreciate Ben & Jerry's pretty uniformly, and they don't eat so many ice-creams that it becomes any budgetary issue worth worrying about.

Regarding the effects of promotional methods, we can review the below 2 plots:



These plots suggest coupons seem to be a bit better than other promotional methods at getting people to spend more.

2. REGRESSION ANALYSIS

We fit the below regression model to detect which variables are statistically significant in affecting the average price paid per ice-cream:

```
glm_model <- train(log(1 + price_paid_per_1) ~
                     log(1 + coupon_value_per_1) + promotion_type +
                     log(1 + household_income) +
                     male_head_employed_full_time +
                     female_head_employed_full_time +
                     male_head_graduated_college +
                     female_head_graduated_college +
                     married + children_under_18 +
                     one_family_house + microwave + dishwasher +
                     garbage_disposal + cable_tv + internet +
                     race + region +
                     size1_descr + flavor_descr + formula_descr,
                   data = copy(benjer),
                   method = "glm")
results <- coef(summary(glm_model$finalModel))[, c("Estimate", "Pr(>|t|)")]
```

```
results
```

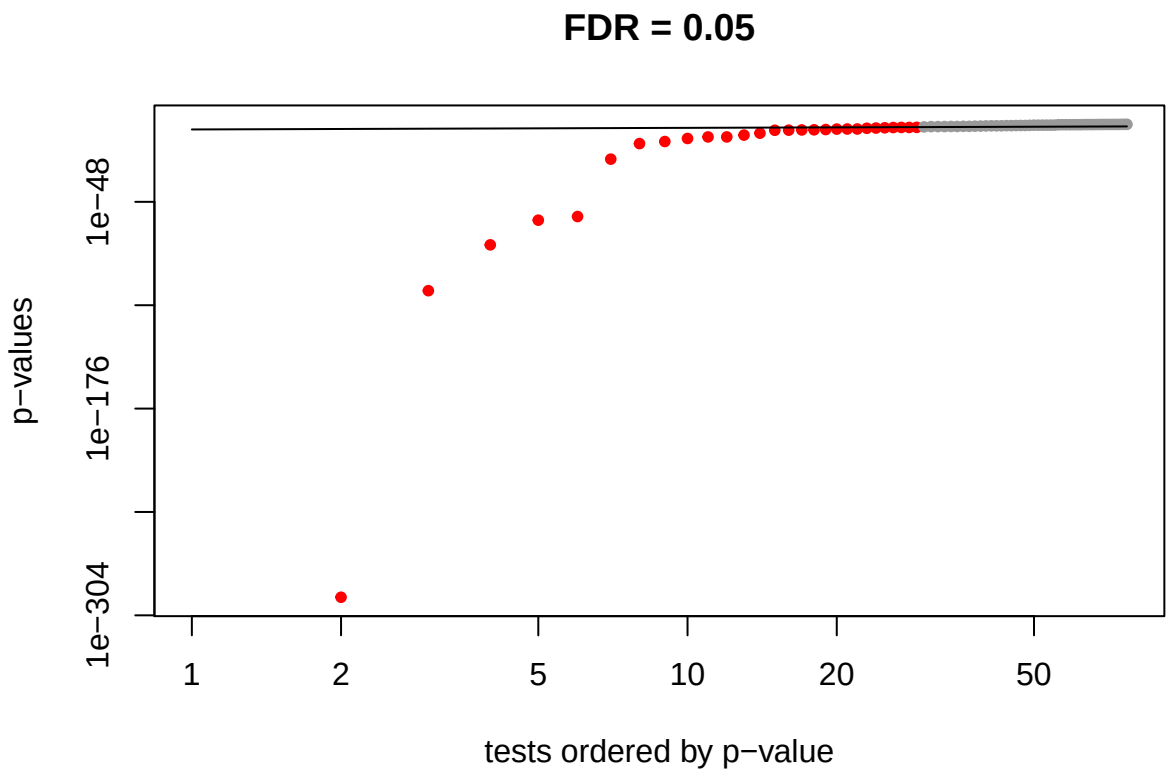
```
##                                                    Estimate      Pr(>|t|)
## (Intercept)                                      1.4557257862  0.000000e+00
## `log(1 + coupon_value_per_1)`                    0.1720052781 9.950772e-104
## promotion_typeStoreFeature                      -0.0372467345  8.253696e-58
## promotion_typeStoreCoupon                       -0.1185147078  4.397835e-60
## promotion_typeManufacturerCoupon                -0.1273030581  2.372099e-75
## promotion_typeOtherDeal                         -0.0486202736  3.295906e-04
## `log(1 + household_income)`                      0.0079879922  2.141922e-11
## male_head_employed_full_timeTRUE                -0.0091022242  2.479662e-04
## female_head_employed_full_timeTRUE              -0.0005628500  7.928190e-01
## male_head_graduated_collegeTRUE                  0.0130843775  1.568533e-08
## female_head_graduated_collegeTRUE                0.0054932179  1.179325e-02
## marriedTRUE                                     -0.0117595542  3.160191e-06
## children_under_18TRUE                           -0.0045893257  5.822705e-02
## one_family_houseTRUE                            -0.0087525611  4.013542e-04
## microwaveTRUE                                   -0.0392625642  2.134065e-07
## dishwasherTRUE                                  -0.0010707617  6.776744e-01
## garbage_disposalTRUE                             0.0039869408  1.910111e-01
## cable_tvTRUE                                     0.0053384374  1.301107e-02
## internetTRUE                                     0.0049560570  8.167987e-02
## raceBlack                                        0.0117302491  3.593544e-03
## raceAsian                                        0.0003531025  9.551246e-01
## raceHispanic                                     0.0098166115  1.529379e-01
## raceOther                                        0.0187509839  1.179598e-02
## regionCentral                                   -0.0222462807  1.153764e-12
## regionSouth                                     -0.0289456366  2.787050e-22
## regionWest                                      -0.0183255696  1.733540e-09
## `size1_descr32.0 MLOZ`                           0.3012740709 1.596420e-293
## `flavor_descrBANANA SPLIT`                      -0.0155587961  4.739063e-02
## `flavor_descrBLACK & TAN`                       -0.0593258590  4.741055e-02
## `flavor_descrBROWNIE BATTER`                     0.0249756680  5.810604e-02
## `flavor_descrBUTTER PECAN`                       0.0168144632  1.199834e-01
## `flavor_descrCAKE BATTER`                       -0.0137582020  1.201757e-01
## flavor_descrCHC                                  0.0156290157  3.254696e-01
## `flavor_descrCHC ALMOND NOUGAT`                  0.0086726994  5.474116e-01
## `flavor_descrCHC CHIP C-DH`                      0.0014621615  8.307394e-01
## `flavor_descrCHC FUDGE BROWNIE`                  0.0004137081  9.497183e-01
## `flavor_descrCHERRY GRCA`                       -0.0033424538  5.787562e-01
## `flavor_descrCHUBBY HUBBY`                       0.0199450695  3.944413e-02
## `flavor_descrCHUNKY MONKEY`                     -0.0097741075  1.481281e-01
## `flavor_descrCINNAMON BUNS`                     -0.0199864790  1.017001e-02
## flavor_descrCOFFEE                               0.0571939817  4.887342e-03
## `flavor_descrCREME BRULEE`                      -0.0230286548  7.030596e-03
## `flavor_descrDOUBLE CHC FUDGE SWR`               0.0332993247  8.210261e-01
## `flavor_descrDUBLIN MUDSLIDE`                    0.0149026328  1.044256e-01
## `flavor_descrFOSSIL FUEL`                       -0.0172583314  3.056275e-01
## `flavor_descrHALF BAKED`                        -0.0075912495  3.108108e-01
## `flavor_descrHEATH CANDY EVERYTHING BUT THE`    -0.0084701252  2.995743e-01
## `flavor_descrHEATH COFFEE CRUNCH`                0.0045603591  4.995620e-01
## `flavor_descrHEATH CRUNCH`                       0.0269262567  1.239844e-03
## `flavor_descrIMAGINE WHIRLED PEACE`             -0.0139473027  7.360472e-02
## `flavor_descrKARAMEL SUTRA`                     -0.0029634573  6.887520e-01
```

```
## `flavor_descrMAGIC BROWNIES`                -0.0116989267  3.125993e-01
## `flavor_descrMINT CHC CHUNK`                 0.0426858054  1.243871e-03
## `flavor_descrNEAPOLITAN DYNAMITE`            0.0068119949  5.654646e-01
## `flavor_descrNEW YORK SUPER FUDGE CHUNK`    -0.0042995994  5.374158e-01
## `flavor_descrOATMEAL COOKIE CHUNK`           0.0050393422  6.739694e-01
## `flavor_descrONE CSK BROWNIE`               -0.0297715297  2.085726e-04
## `flavor_descrOXFORD MINT CHC COOKIE`         0.0014818590  8.771014e-01
## `flavor_descrPB CUP`                        -0.0008500305  9.055796e-01
## `flavor_descrPB TRUFFLE`                    -0.0585523590  6.907275e-01
## `flavor_descrPHISH FOOD`                     0.0024505121  7.341459e-01
## `flavor_descrPISTACHIO PISTACHIO`            0.0018774947  8.008280e-01
## `flavor_descrPUMPKIN CSK`                   -0.0752272871  1.531570e-08
## `flavor_descrRSP CHC CHUNK`                  0.0112168828  5.740771e-01
## flavor_descrSMORES                           0.0399613700  5.861755e-04
## flavor_descrSTR                              0.0701351263  1.164244e-01
## `flavor_descrSTR CSK`                       -0.0171835377  3.622479e-02
## `flavor_descrSTRAWBERRIES & CREAM`          -0.0119540401  7.771291e-01
## `flavor_descrSWEET CREAM & COOKIES`          0.0465518139  1.971414e-01
## `flavor_descrTRIPLE CARAMEL CHUNK`          -0.0038294284  8.182848e-01
## `flavor_descrTURTLE SOUP`                   -0.0233848959  4.141418e-02
## flavor_descrVAN                              0.0268961557  1.109181e-03
## `flavor_descrVAN CARAMEL FUDGE`             -0.0120134690  2.348067e-01
## `flavor_descrVERMONTY PYTHON`                0.0082803956  5.448689e-01
## `flavor_descrW-N-C-P-C`                     -0.0079144440  2.935819e-01
## `flavor_descrWHITE RUSSIAN`                  0.2003418576  1.734875e-01
## formula_descrREGULAR                        -0.0217240574  2.555692e-02
```

The only clear, interpretable result from this analysis is that coupons are statistically significant. For other variables, even though we see some pretty small p-values, the signs of the coefficents do not lend themselves for very meaningful interpretation.

3. MANAGING THE FALSE DISCOVERY RATE (FDR)

```
pvals <- results[, 2]
source("fdr.R")
alpha = fdr_cut(pvals, 0.05, plotit=TRUE)
```



With the above model, in order to guarantee an expected False Discovery Rate (FDR) of at most 5%, we need to use the cut-off threshold of **0.0130111** for the p-values. We need to pay attention to the FDR in models that have many covariates such as this one to mitigate the probability of deciding that a certain variable is statistically significant when it is in fact not.

With this cut-off, the variables that are statistically significant are:

```
results[pvals < alpha, ]
```

```
##                                    Estimate      Pr(>|t|)
## (Intercept)                       1.455725786  0.000000e+00
## `log(1 + coupon_value_per_1)`     0.172005278 9.950772e-104
## promotion_typeStoreFeature       -0.037246734  8.253696e-58
## promotion_typeStoreCoupon        -0.118514708  4.397835e-60
## promotion_typeManufacturerCoupon -0.127303058  2.372099e-75
## promotion_typeOtherDeal          -0.048620274  3.295906e-04
## `log(1 + household_income)`       0.007987992  2.141922e-11
## male_head_employed_full_timeTRUE -0.009102224  2.479662e-04
## male_head_graduated_collegeTRUE   0.013084378  1.568533e-08
## female_head_graduated_collegeTRUE 0.005493218  1.179325e-02
## marriedTRUE                      -0.011759554  3.160191e-06
## one_family_houseTRUE             -0.008752561  4.013542e-04
## microwaveTRUE                    -0.039262564  2.134065e-07
## raceBlack                         0.011730249  3.593544e-03
## raceOther                         0.018750984  1.179598e-02
## regionCentral                    -0.022246281  1.153764e-12
```

```
## regionSouth                       -0.028945637  2.787050e-22
## regionWest                        -0.018325570  1.733540e-09
## `size1_descr32.0 MLOZ`             0.301274071 1.596420e-293
## `flavor_descrCINNAMON BUNS`       -0.019986479  1.017001e-02
## flavor_descrCOFFEE                 0.057193982  4.887342e-03
## `flavor_descrCREME BRULEE`        -0.023028655  7.030596e-03
## `flavor_descrHEATH CRUNCH`         0.026926257  1.239844e-03
## `flavor_descrMINT CHC CHUNK`       0.042685805  1.243871e-03
## `flavor_descrONE CSK BROWNIE`     -0.029771530  2.085726e-04
## `flavor_descrPUMPKIN CSK`         -0.075227287  1.531570e-08
## flavor_descrSMORES                 0.039961370  5.861755e-04
## flavor_descrVAN                    0.026896156  1.109181e-03
```

The few *flavor_decr* variables remaining in this list are highly suspicious and don't look like true discoveries, given that most other *flavor_decr* variables failed to satisfy the significance test with the above threshold. The purpose of controlling the FDR is exactly to minimize the number of false discoveries like these.

Even within the subset of variables that are deemed to be statistically significant with the above threshold, I feel many of them are false discoveries and am hesitant to read too much into them.

Overall, I think the following are likely to be true discoveries:

- The use of coupons;
- Household income; and
- Whether or not the family heads are college graduates