Walmart Sales Analysis

Holden Qiu

2020/05/21

Contents

1 Quick Look

As a first step let's have a quick look of the data sets using the `head`, `summary`, and `glimpse` tools where appropriate.

1.1 Training sales data

Here are the first 10 columns and rows of the our training sales data:

id	item_id	dept_id	cat_id	store_id	state_id	d_1	d_
HOBBIES_1_001_CA_1_validation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	0	
HOBBIES_1_002_CA_1_validation	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	0	
HOBBIES_1_003_CA_1_validation	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	0	
HOBBIES_1_004_CA_1_validation	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	0	
HOBBIES_1_005_CA_1_validation	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	0	
HOBBIES_1_006_CA_1_validation	HOBBIES_1_006	HOBBIES_1	HOBBIES	CA_1	CA	0	
HOBBIES_1_007_CA_1_validation	HOBBIES_1_007	HOBBIES_1	HOBBIES	CA_1	CA	0	
HOBBIES_1_008_CA_1_validation	HOBBIES_1_008	HOBBIES_1	HOBBIES	CA_1	CA	12	
HOBBIES_1_009_CA_1_validation	HOBBIES_1_009	HOBBIES_1	HOBBIES	CA_1	CA	2	
HOBBIES_1_010_CA_1_validation	HOBBIES_1_010	HOBBIES_1	HOBBIES	CA_1	CA	0	

We find:

- There is one column each for the IDs of item, department, category, store, and state; plus a general ID that is a combination of the other IDs plus a flag for validation.
- The sales per date are encoded as columns starting with the prefix `d_`. Those are the number of units sold per day (not the total amount of dollars).
- We already see that there are quite a lot of zero values.

This data set has too many columns and rows to display them all:

```
## [1]  1919 30490
```

1.2 Missing values & zero values

There are no missing values in our sales training data:

```
## [1] 0
```

However, there are a lot of zero values, here we plot the distribution of zero percentages among all time series:

This means that only a minority of time series have less than 50% of zero values. The peak is rather close to 100%

2 Visual Overview: Interactive time series plots

We will start our visual exploration by investigating a number of time series plots on different aggregation levels. Here is a short helper function to transform our wide data into a long format with a `dates` column in date format:

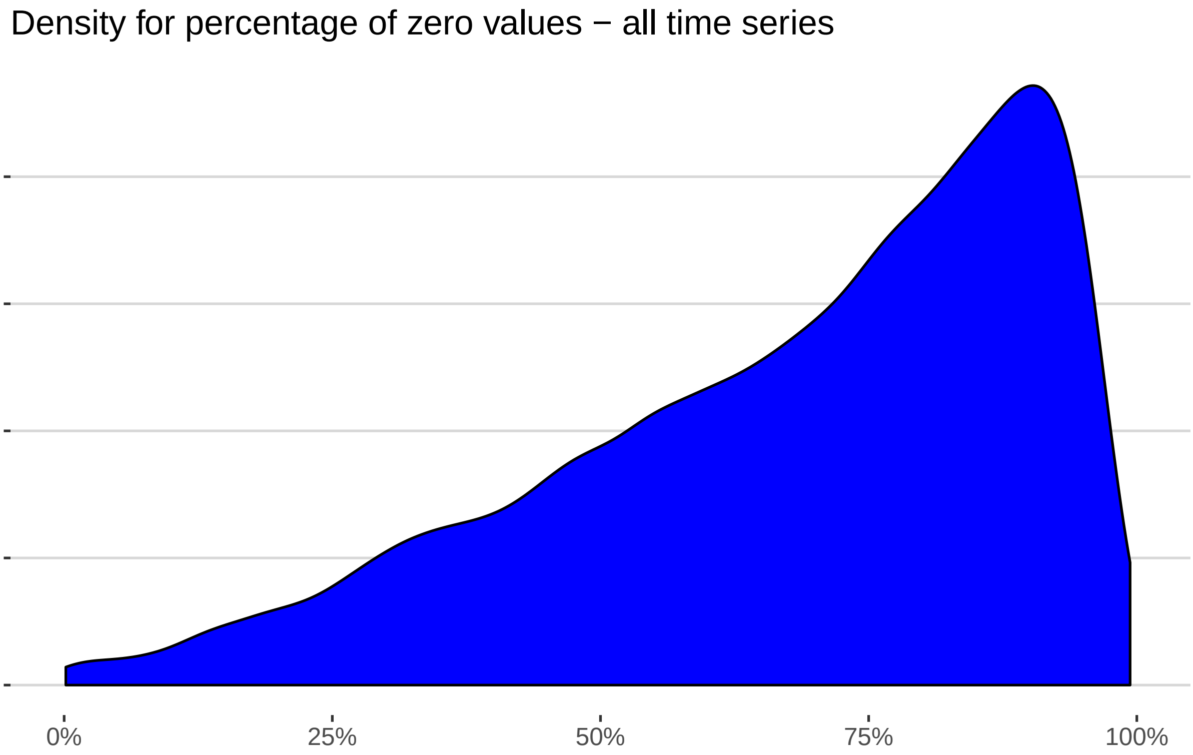

Figure 1: Fig. 1

2.1 Individual item-level time series - random sample

Here we will sample 50 random time series from our training sample to get an idea about the data.

In the following plot, you can select the id of a time series (which is a concatenation of store and department) to display only the graphs that you're interested in.

Currently, I don't see a way to avoid having all time series plotted at the beginning. Select 1 from the input field to begin:

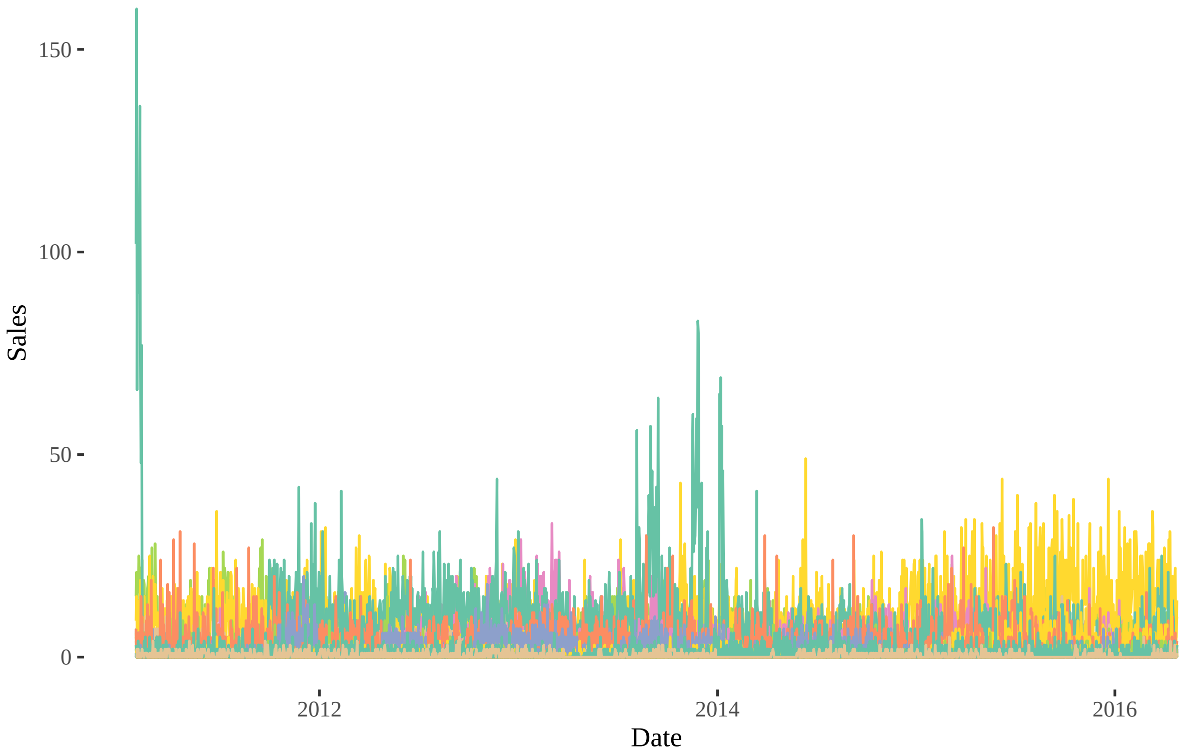

Figure 2: Fig. 2

```
## NULL
```

We find:

- Most time series have pretty low daily count statistics, alongside the large percentage of zero numbers we already noticed. On the one hand, this suggests that spikes are not going to be overly pronounced. But it also indicates that accurate forecasts will have to deal with quite a lot of noise.
- Some of our sample time series start in the middle of the time range, and some have long gaps in between. This is an additional challenge.

2.2 All aggregate sales

After peeking at some of the individual time series and finding a lot of zero values, we will now do some aggregation to get some decent statistics.

First off, here we plot the aggregate time series over all items, stores, categories, departments and sales. This is an interactive plot and you can use the usual plotly tools (upper right corner) to zoom, pan, and scale the view.

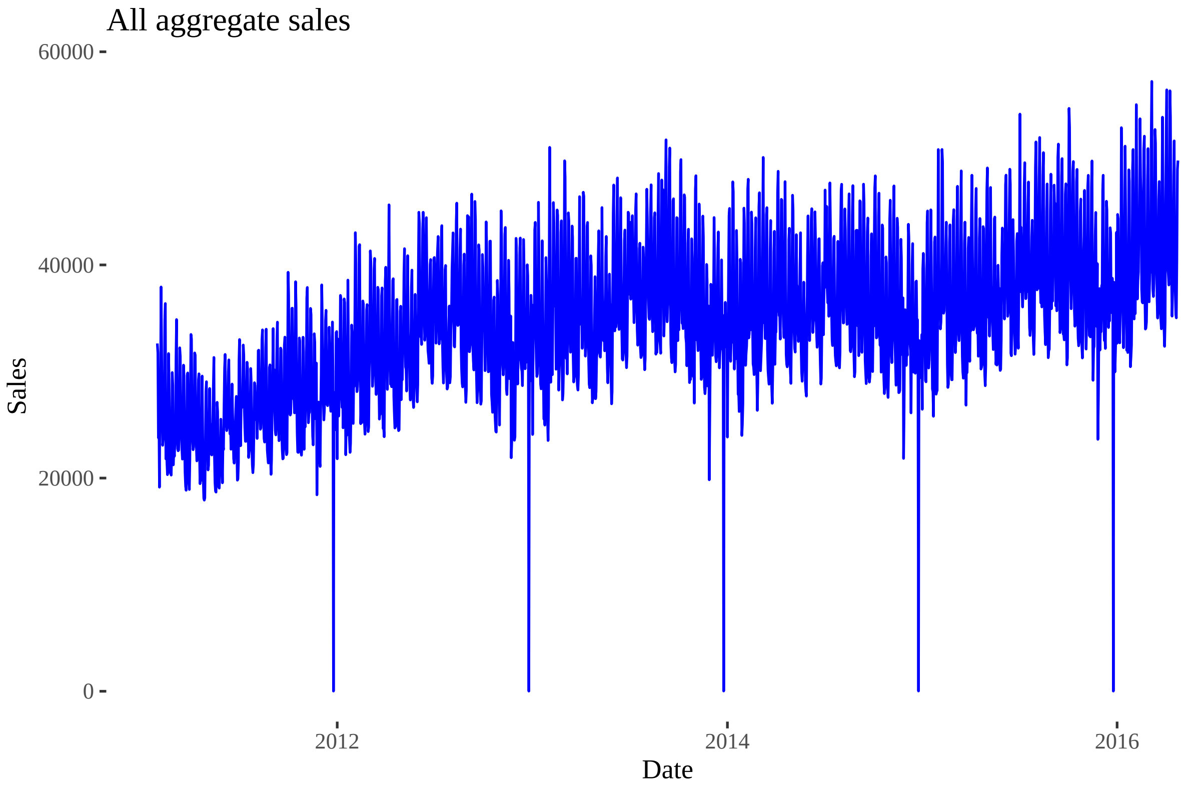

Figure 3: Fig. 3

We find:

- The sales are generally going up, which I suppose is good news for Walmart. We can make out some yearly seasonality, and a dip at Christmas, which is the only day of the year when the stores are closed.
- Zooming in, we can see strong weekly seasonality plus possibly some additional overlaying patterns with shorter periods than yearly.
- The most recent 2016 sales numbers appear to grow a bit faster than in previous years.

2.3 Sales per State

To get a bit more out of these big picture views we will look at the sales per state on a monthly aggregate level. This is another interactive `ggplotly` graph:

We find:

- California (CA) sells more items in general, while Wisconsin (WI) was slowly catching up to Texas (TX) and eventually surpassed it in the last months of our training data.
- CA has pronounced dips in 2013 and 2015 that appear to be present in the other states as well, just less severe. These dips and peaks don't appear to always occur (see 2012) but they might primarily reflect the yearly seasonality we noticed already.

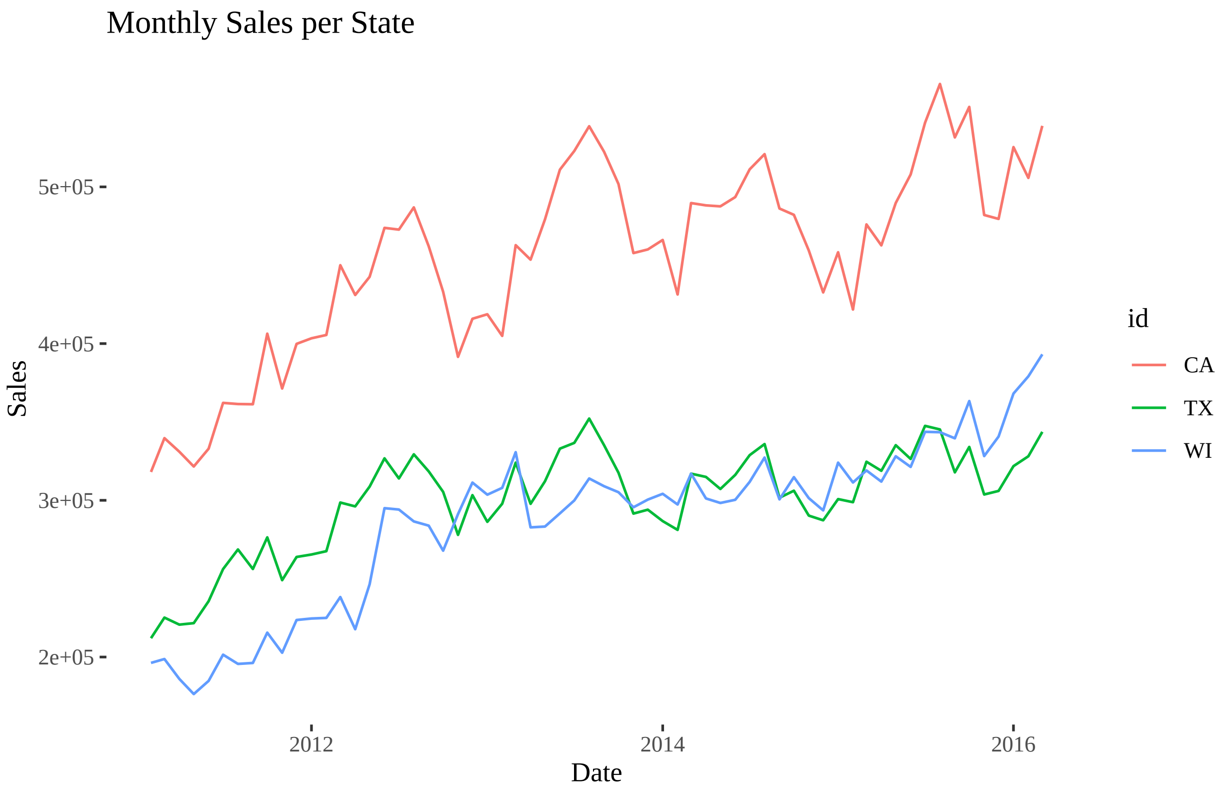

Figure 4: Fig. 4

2.4 Sales per Store & Category

There are 10 stores, 4 in CA and 3 each in TX and WI, and 3 categories: Foods, Hobbies, and Household. Here we're switching up our visualisation strategy a bit by including all of those in the same, non-interactive layout. We will again use monthly aggregates to keep the plots clean.

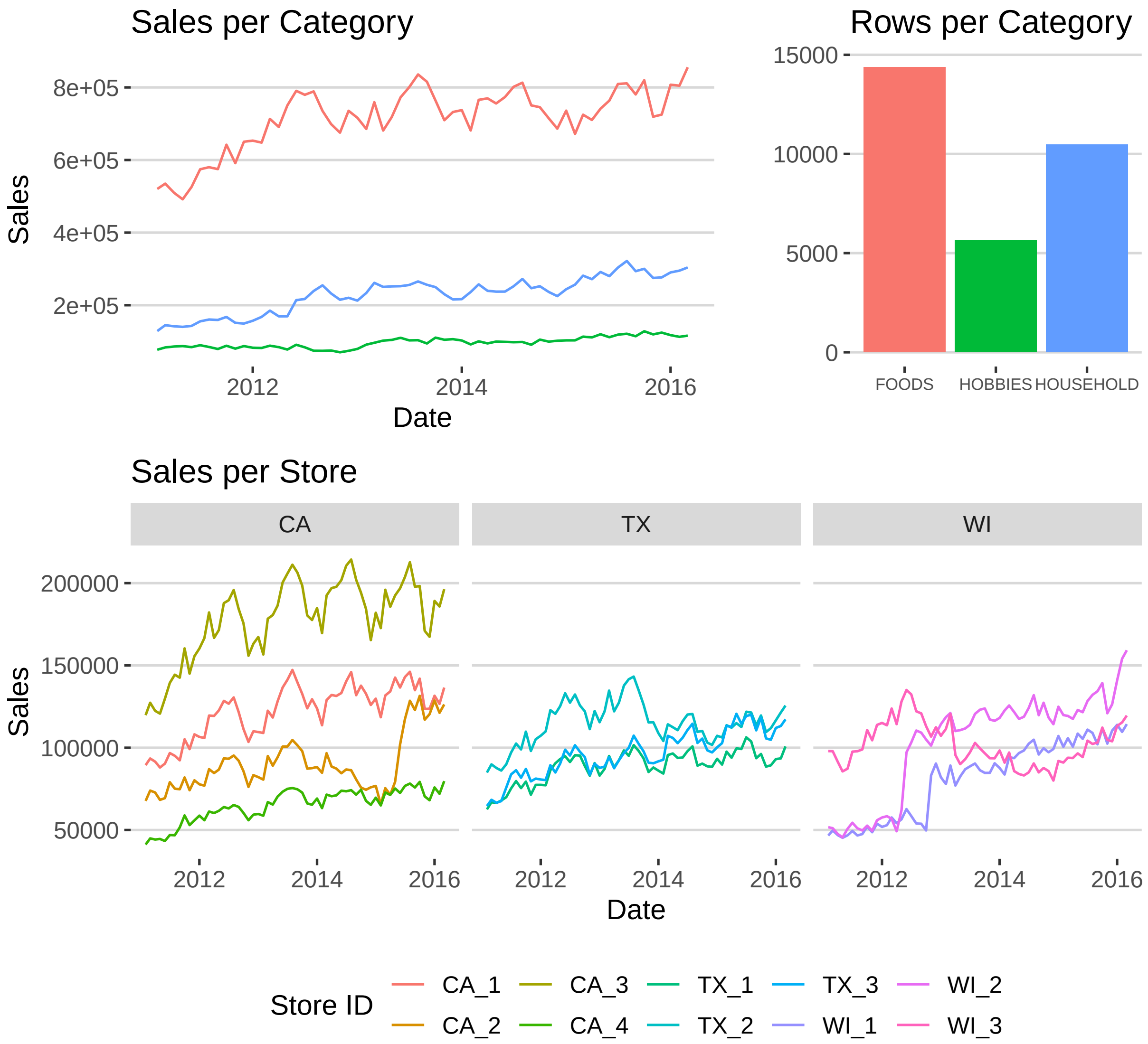

Figure 5: Fig. 5

We find:

- "Foods" are the most common category, followed by "Household" which is still quite a bit above "Hobbies". The number of "Household" rows is closer to the number of "Foods" rows than the corresponding sales figures, indicating that more "Foods" units are sold than "Household" ones.
- In terms of stores, we see that the TX stores are quite close together in sales; with "TX_3" rising from the levels of "TX_1" to the level of "TX_2" over the time of our training data. The WI stores "WI_1" and "WI_2" show a curious jump in sales in 2012, while "WI_3" shows a long dip over several year.

- The CA stores are relatively well separated in store volume. Note “CA_2”, which declines to the “CA_4” level in 2015, only to recover and jump up to “CA_1” sales later in the year.

2.5 Sales per Department

Our data has 7 departments, 3 for “FOODS” and 2 each for “HOBBIES” and “HOUSEHOLD”. Together with the 3 states those are 21 levels. Let’s see whether we can fit them all in a single facet grid plot.

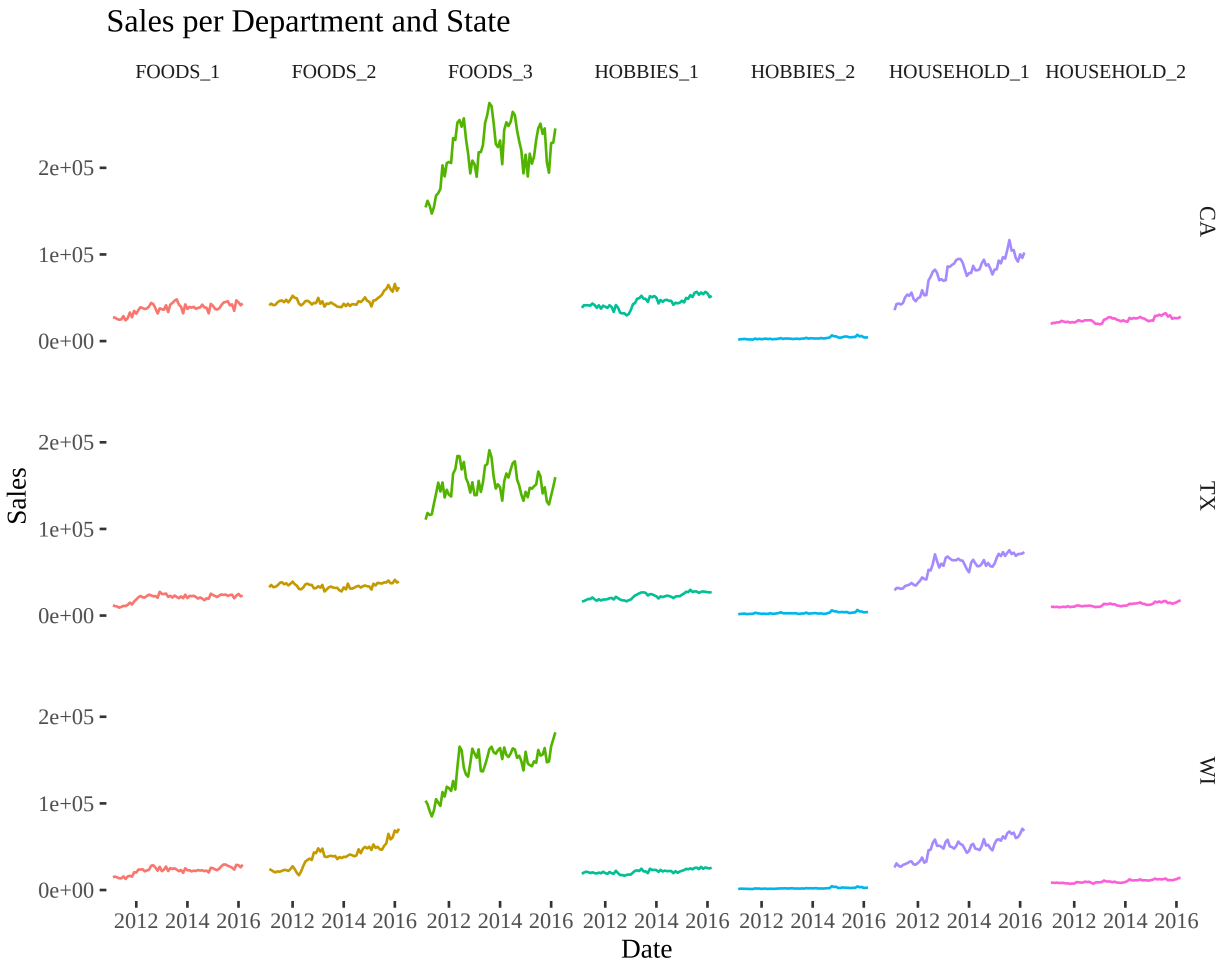

Figure 6: Fig. 5

We find:

- “FOODS_3” is clearly driving the majority of “FOODS” category sales in all states. “FOODS_2” is picking up a bit towards the end of the time range, especially in “WI”.
- Similarly, “HOUSEHOLD_1” is clearly outselling “HOUSEHOLD_2”. “HOBBIES_1” is on a higher average sales level than “HOBBIES_2”, but both are not showing much development over time.

2.6 Seasonalities - global

Moving on from the time series views, at least for the moment, we are changing up our visuals to study seasonalities. Here is a heat map that combines the weekly and yearly seasonalities.

Because of the general increasing trend in sales, we're not looking at absolute sales values. Instead, we aim to model this trend using a smoothed (LOESS) fit which we then subtract from the data. The heatmap shows the relative changes. Note, that I removed the Christmas dips because they would be distracting for the purpose of this plot:

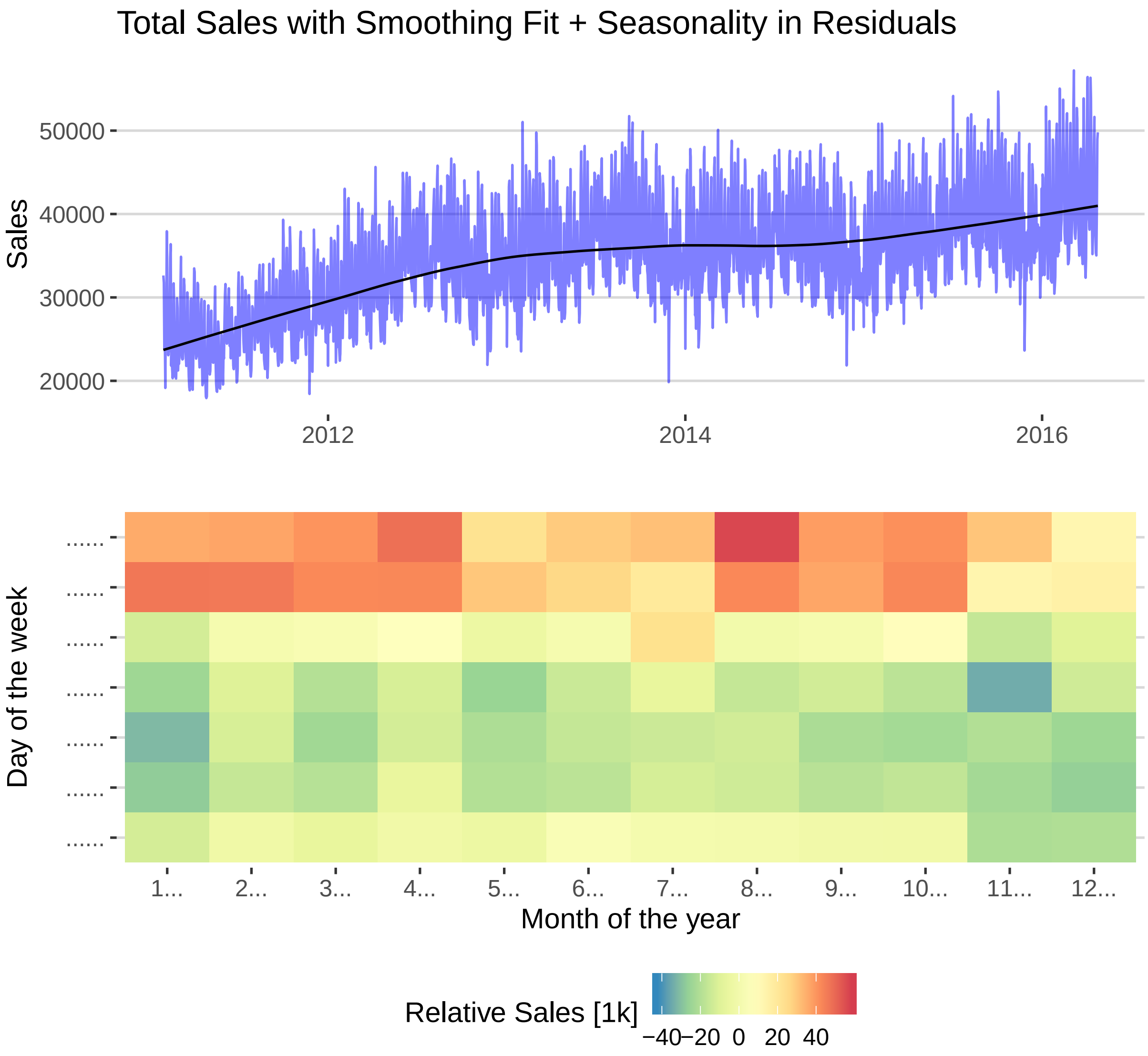

Figure 7: Fig. 6

We find:

- The weekly pattern is strong, with Sat and Sun standing out prominently. Also Monday seems to benefit a bit from the weekend effect.

- The months of Nov and Dec show clear dips, while the summer months May, Jun, and Jul suggest a milder secondary dip. Certain holidays, like the 4th of July, might somewhat influence these patterns; but over 5 years they should average out reasonably well.
- I already tweaked the smoothing fit parameters a bit, but they could probably be optimised further. Still, it looks quite good for the first try.

Let's stay with the seasonalities for a little while longer and go one or two levels deeper.

Next, we'll look at the weekly and monthly patterns on a state level. We're using the same smoothing approach which is shown in the upper panels for reference. Here, I will scale each individual time series by its global mean to allow for a better comparison between the three states.

The colours are the same throughout the layout; see the upper panels for reference:

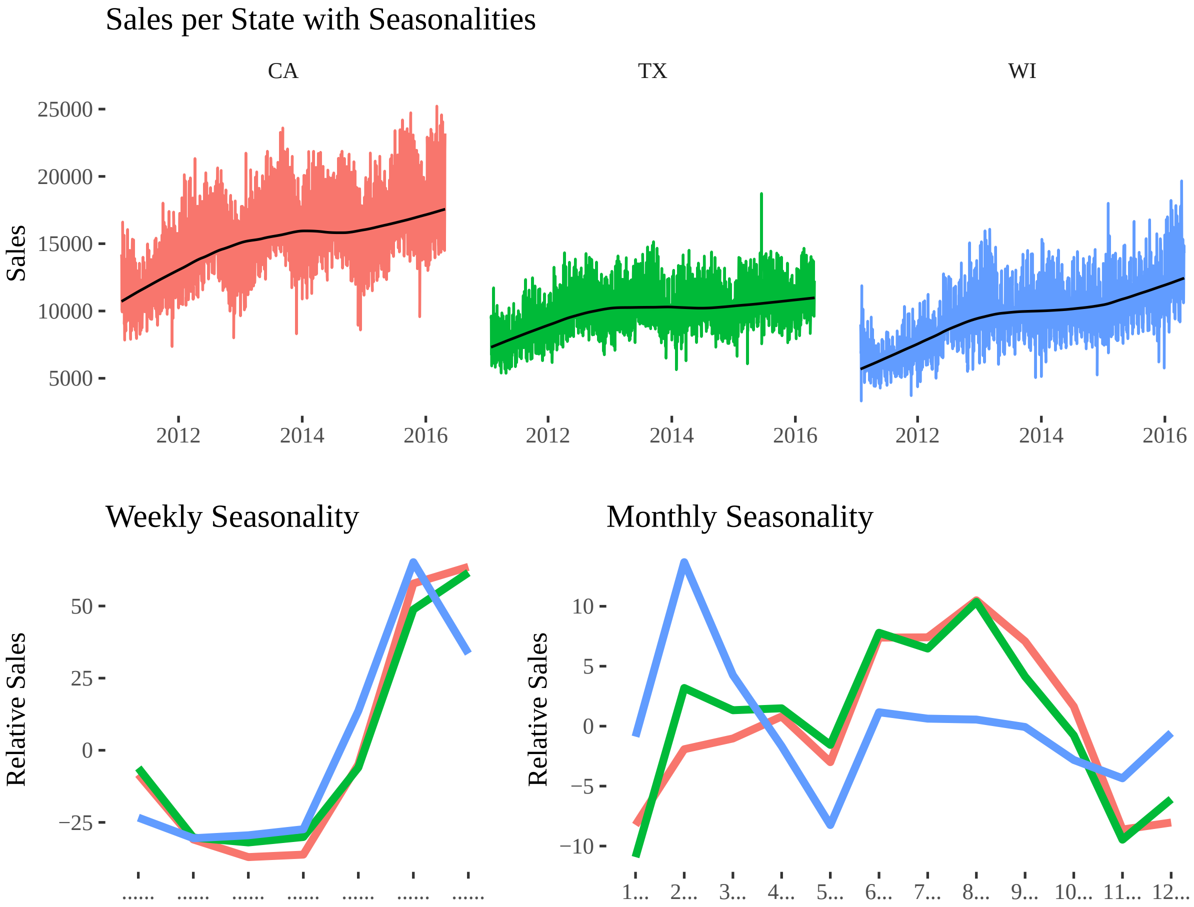

Figure 8: Fig. 7

We find:

- After scaling, the weekday vs weekend pattern is very similar for all 3 states, except for an interesting downturn in Sunday sales in WI.

- The monthly seasonalities are indeed complex. There is a dip in the winter months and a second, generally shallower dip around May. WI is again the odd state out: it sells notably less in the summer compared to TX and especially CA; so much so that the Feb/Aug ratio is inverted for WI vs CA/TX.

The logical next step is to look at the Seasonalities per Category, since you wouldn't necessarily expect food and household item shopping to follow the same patterns. I will also break up this view by state, following the insights we just saw above.

Again, we are subtracting the trend and scaling by the global mean:

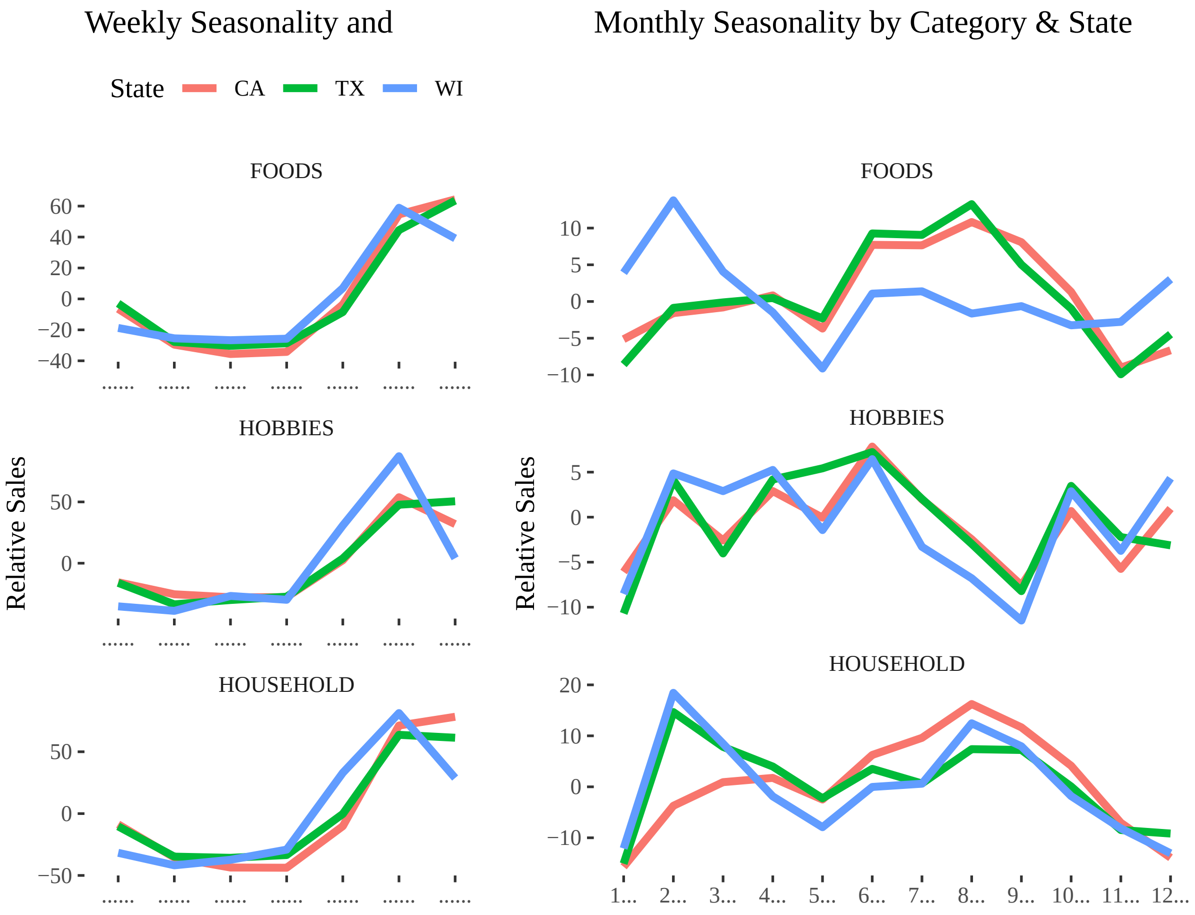

Figure 9: Fig. 8

We find:

- The weekly patterns for the "foods" category are very close for all 3 states, and WI shows some deviations for "hobbies" and "household". We also see the Wisconsin's characteristic Sunday dip for all 3 categories.
- The monthly patterns show some interesting signatures: For "foods", CA and TX are pretty close but WI shows that inverted summer/winter ratio. In contrast, the 3 states are much more similar to each other in the "hobbies" category. And when it comes to "household" items, CA doesn't seem to sell as much of them during the first 3 months of the year but slightly more in the summer; compared to WI and TX.

Before we look at the additional explanatory variables, here is a visual representation of the split between training data vs validation data (public leaderboard) vs evaluation data (eventual private leaderboard):

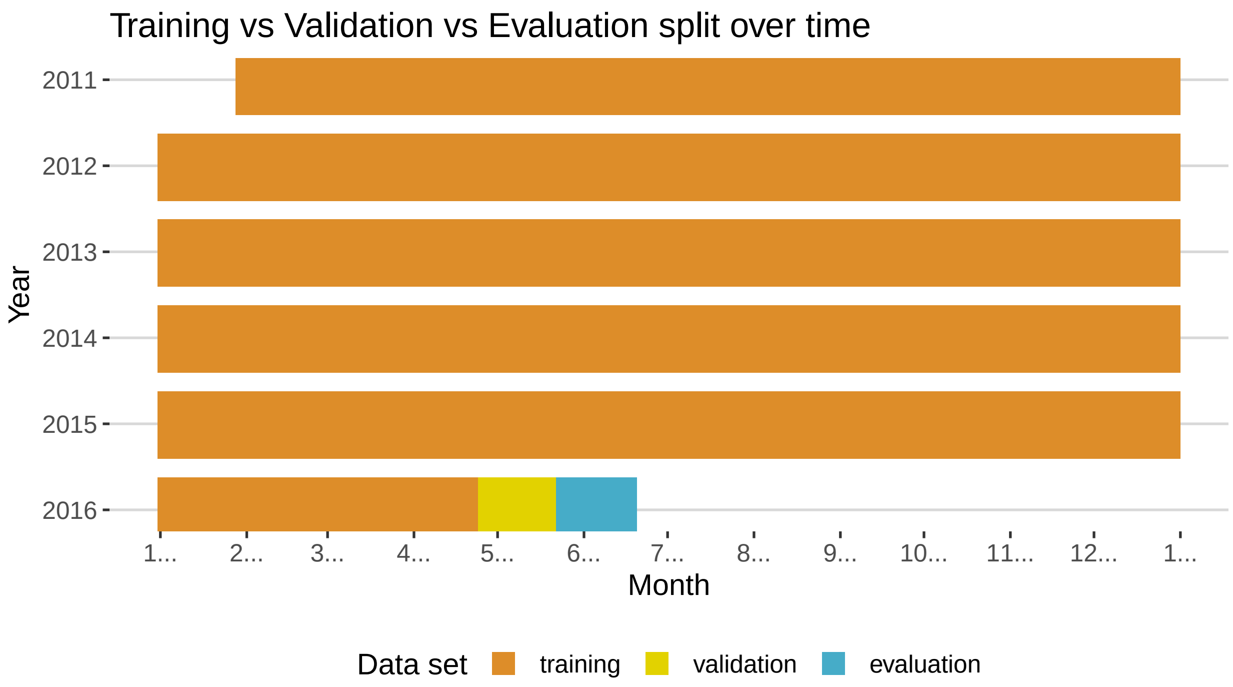

Figure 10: Fig. 9

- This shows the 5+ years of training data together with the 28 days each of validation and evaluation time range.
- The training range goes from 2011-01-29 to 2016-04-24. The validation range spans the following 28 days from 2016-04-25 to 2016-05-22. This is what we are predicting for the initial public leaderboard. Eventually, we will be given the ground truth for this period to train our model for the final evaluation.
- The evaluation range goes from 2016-05-23 to 2016-06-19; another 28 days. This is what the ultimate scoring will be based on.

3 Explanatory Variables: Prices and Calendar

This section will focus on the additional explanatory variables we've been given: item prices and calendar events. In terms of calendar features, we have already used some parameters like day-of-week or month, derived from the date, in the previous section. After studying the basic properties of these datasets, we will also connect them to the time series data.

3.1 Calendar

We need some calendar features for the item price table, so let's start with that one.

In the quick view in section 3.3 we see that the `calendar` data frame contains basic features like day-of-week (as character column `weekday` and as integer column `wday`), `month`, `year`, and of course `date`. Alongside the `date` there is also a `d` column which links the date to the column names in the training data. (Our time series extraction function has the starting date as a hard-coded value, but you can also convert from column name to date using the calendar file.)

The other features deal with events and food stamps:

- If you look at section 3.3. you'll see that the columns `event_name_2` and `event_type_2` only contain 5 values that are not NAs, so we're ignoring them here and only focus on the `event_*_1` features.
- The acronym SNAP stands for Supplemental Nutrition Assistance Program. The following is copied from their website: "The Supplemental Nutrition Assistance Program (SNAP) is the largest federal nutrition assistance program. SNAP provides benefits to eligible low-income individuals and families via an Electronic Benefits Transfer card. This card can be used like a debit card to purchase eligible food in authorized retail food stores."

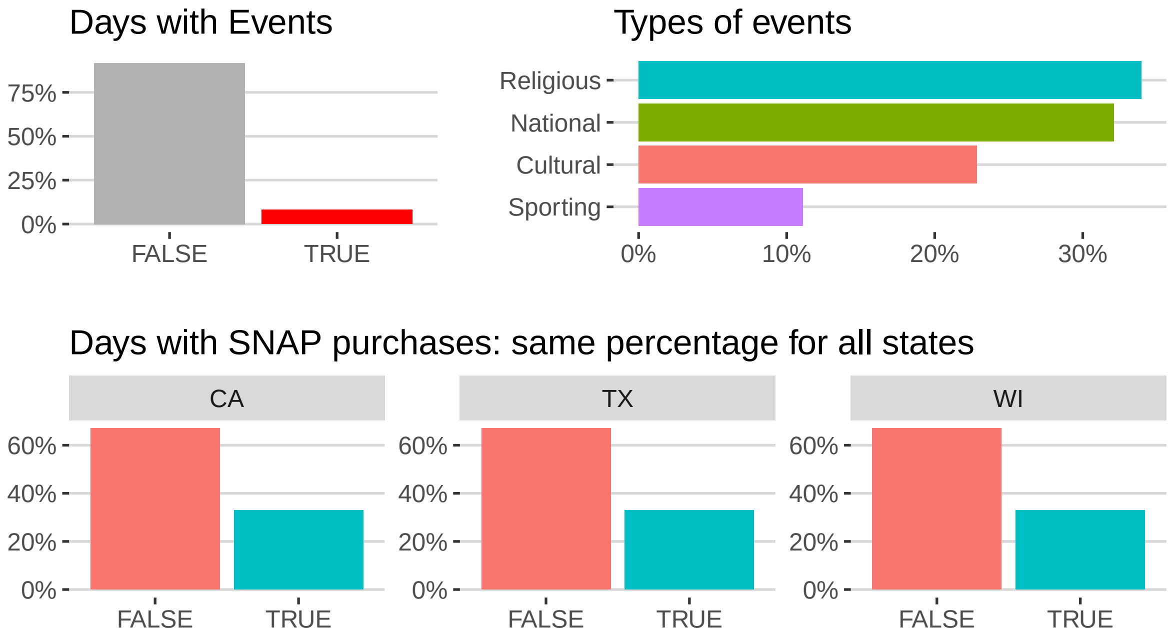

Figure 11: Fig. 10

We find:

- In our calendar coverage (i.e. training + validation + evaluation range) about 8% of days have a special event. Of these events, about 1/3 are Religious (e.g. Orthodox Christmas) and 1/3 are National Holidays (e.g. Independence Day). The remaining third is again split into 2/3 Cultural (e.g. Valentines Day) and 1/3 Sporting events (e.g. SuperBowl).
- Looking at the percentage of days where purchases with SNAP food stamps are allowed in Walmart stores, we find that it is the exact same for each of the 3 states: 650 days or 33%. This is noteworthy.

Let's stay with SNAP for a little bit and look at a different style of visual: a calendar view. Here, I'm showing the days for each month in the shape of days-of-week as rows and weeks-of-month as columns; essentially the view a calendar is displayed in a Year view. Then I colour the SNAP days orange and show the same graph for each state:

```
## tk_augment_timeseries_signature(): Using the following .date_var variable: date
```

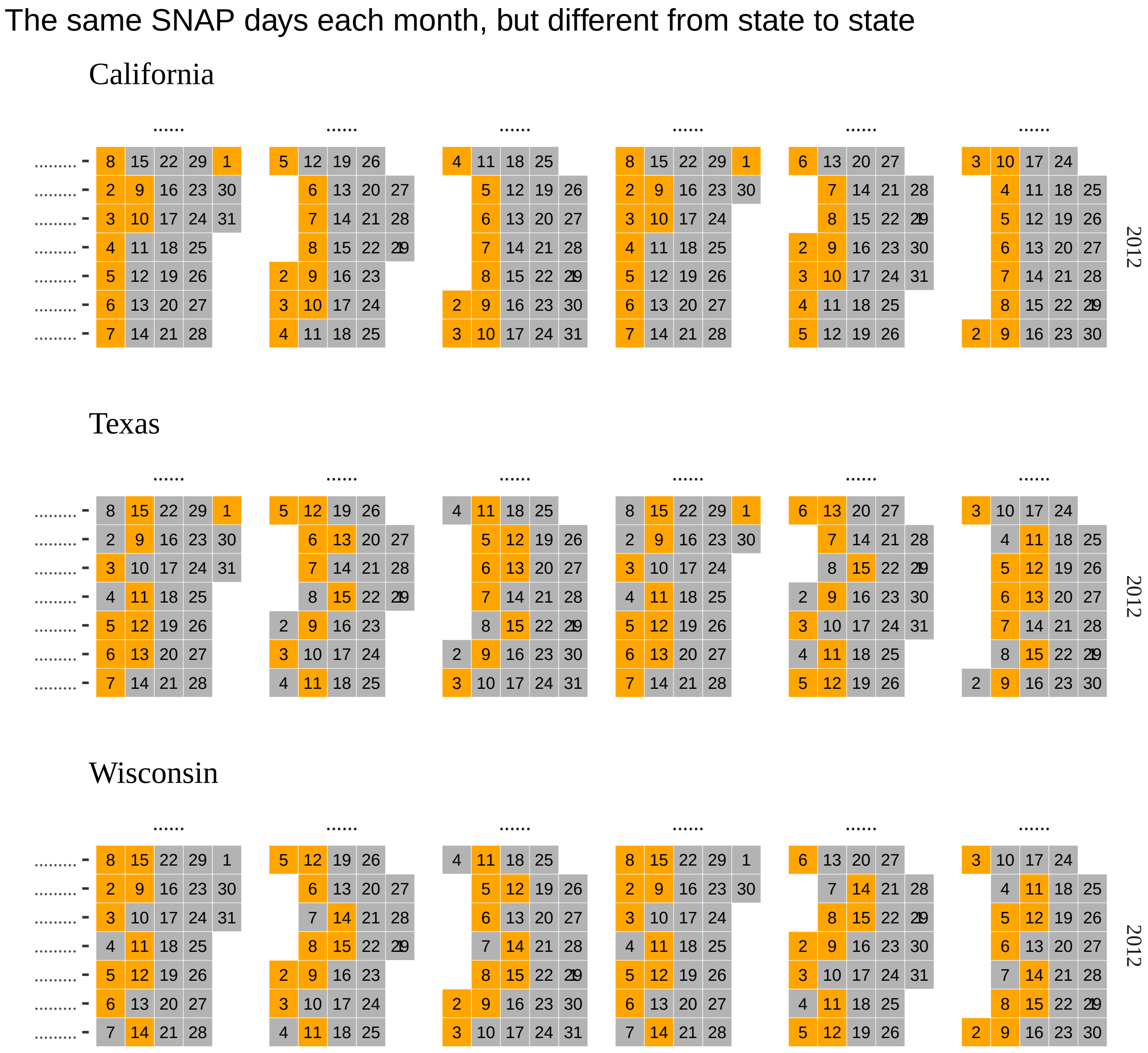

Figure 12: Fig. 11

These are just the first six month of 2012, to keep the plot manageable, but the pattern is always the same:

- SNAP days are always the same days of the month for each month. There are always 10 of them, and the specific days are different from state to state: days 1-10 for CA, days 1-15 without 2, 4, 8, 10, 14 for TX, and days 2-15 without 4, 7, 10, 13 for WI.
- The SNAP days for these 3 states also all happen in the first half of each month, no later than the 15th. This certainly helps to measure their impact and make predictions more robust.

3.2 Item Prices

We have item price information for each item ID, which includes the `category` and `department` IDs, and its `store` ID, which includes the `state` ID. Let's look at some average overviews first.

Here is a facet grid with overlapping density plots for price distributions within the groups of `category`, `department`, and `state`. Note the logarithmic scale on the x-axes:

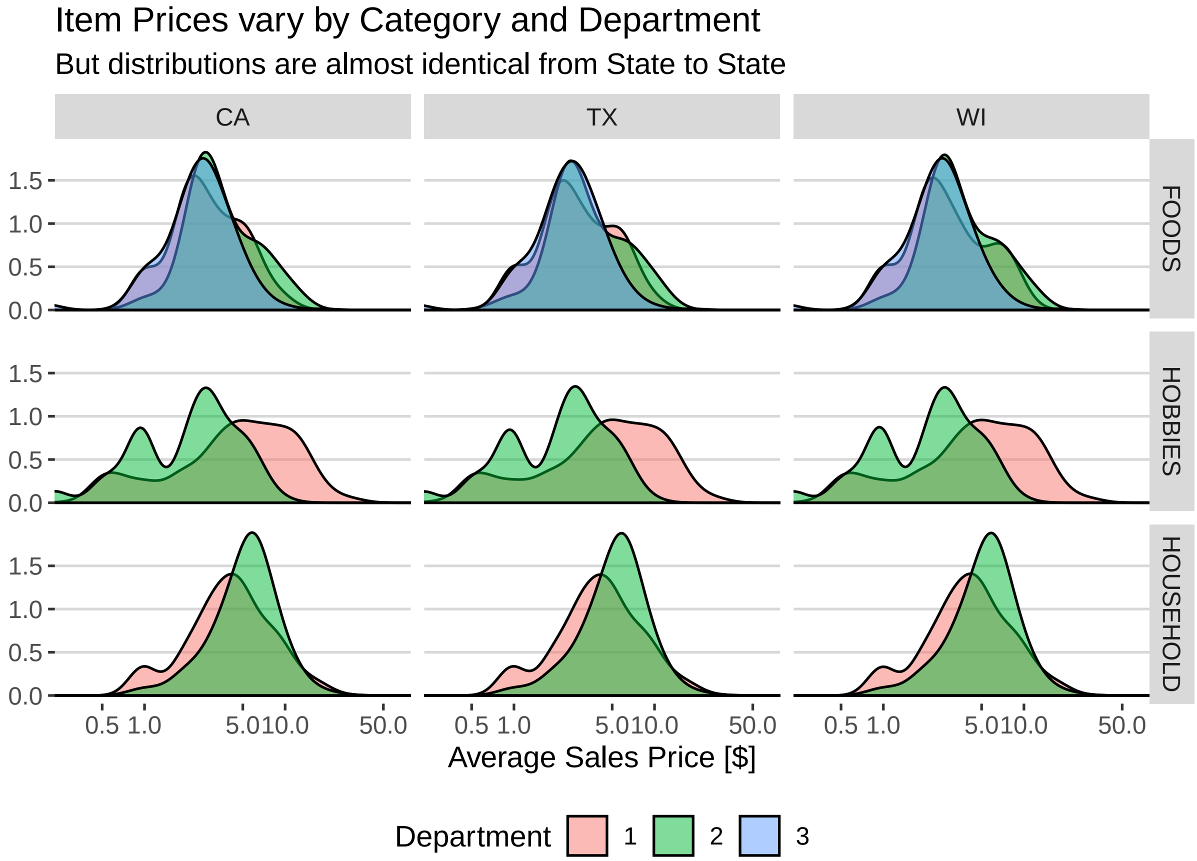

Figure 13: Fig. 12

We find:

- First of all, the distributions are almost identical between the 3 states. There are some minute differences in the "FOODS" category, but this might be due the smoothing bandwidth size. For all practical purposes, I think that we can treat the price distributions as equal.
- There are notable differences between the categories: FOODs are on average cheaper than HOUSEHOLD items. And HOBBIES items span a wider range of prices than the other two; even suggesting a second peak at lower prices.
- Within the categories, we find significant differences, too:
  - Among the three food categories department 3 (i.e. "FOODS_3") does not contain a high-price tail.

  - The HOBBIES category is the most diverse one, with both departments having quite broad distributions but "HOBBIES_1" accounting for almost all of the items above $10. "HOBBIES_2" has a bimodal structure.
  - The HOUSEHOLD price distributions are quite similar, but "HOUSEHOLD_2" peaks at clearly higher prices than "HOUSEHOLD_1".

Prices are provided as weekly averages. The week ID `wm_yr_wk` can be linked to dates (and sales) using the `calendar` column of the same name.

Here we do just that to extract the item price distributions per category and department for each of the year from 2011 to 2016. We use ridgeline plots provided by the ggridges package to produce stacks of overlapping density graphs:

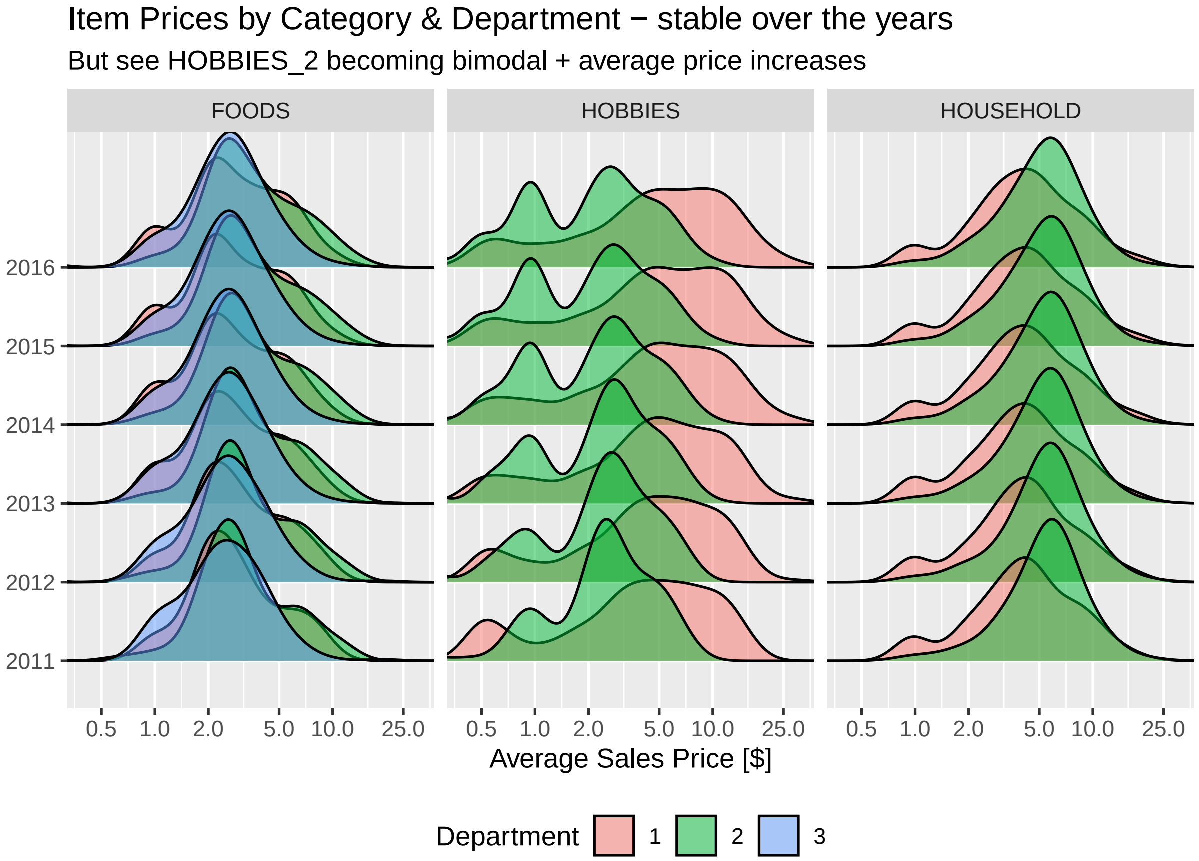

Figure 14: Fig. 13

We find:

- Overall, the price distributions are pretty stable over the years, with only slight increases that are likely due to inflation. For this plot, I've left the standard grey backgrounds so that you can use the vertical grid lines to better see the shifts to the right (remember that the scale is logarithmic). This is probably best visible in HOBBIES_1.
- An interesting evolution is visible in HOBBIES_2, which over time becomes much more bimodal: the second peak at $1 is increasing in importance until it almost reaches the level of the main peak just

above 2 dollars. At the same time the small secondary peak at half a dollar in HOBBIES_1 becomes more flat after 2012.

- The HOUSEHOLD departments are stable. FOODS shows small changes like the relative growth of the $1 peak of FOODS_1.

3.3 Connection to time series data

Now that we have an idea about the properties of our explanatory variables, let's see how they could influence our time series data.

Here, we're looking at time series properties on an aggregate level. I plan for the next section to contain views of selected individual time series.

First off, this is a comprehensive view of sales volume during days with special events vs non-events for our 3 categories food, hobbies, and household. I'm showing the daily time series in the background, but it's more instructive to look at the smoothed representations. I'm doing the smoothing per category and also globally (dashed line). This global fit will be used to compute the relative sales for the two bottom panels. In the lower right panel I only look at days with events and show the median sales for the 4 different types of events. The two bottom panels share the same y-axis labels between them.

Take your time to digest this view:

We find:

- For FOODS the smoothed lines of event vs non-event sales are pretty similar, while for HOBBIES the red event line is consistently below the non-events and for HOUSEHOLD the same is true after 2013. (This is a curious detail, that before 2013 there was no real difference between those sales.)

- This impression is confirmed by looking at the boxplots in the lower left panel, which show the relative sales after subtracting the global smoothing fit. The FOODS sales are pretty comparable between events and non-events, while the event sales for HOUSEHOLD and especially HOBBIES are notably below the non-event level.

- In the lower right panel we break out the events by type and look at the medians of the relative sales. The first thing we see is that FOODS sales are notably higher during "Sporting" events. This makes sense, given the food culture associated with big events like the Superbowl. FOODs also have slightly positive net sales during "Cultural" events.

- In general, "National" and "Religious" events both lead to relative decline in sales volume. "National" events are more depressing for the HOBBIES category, while the other two categories are slightly more affected by "Religious" events. HOBBIES also sees lower sales from "Cultural" events, while for FOODS and HOUSEHOLD the differences are smaller. "Sporting" has a minor impact on HOUSEHOLD and HOBBIES.

We've spend quite a bit of time on designing this dashboard-like overview, so let's get some mileage out of it. Here's the equivalent plot for the 3 states:

We find:

- Special events slightly outsell non-event days in TX before 2014; afterwards they are similar. CA and WI also show a drop around the same time, but here it's from similar sales to lower sales. This seems to be a common pattern that starts in 2013.

- As a result the events boxplot for WI is notably shifted toward lower values, while in TX events push the sales figures to somewhat higher values. CA is roughly the same for events vs non-events. Note, that this is the global picture which will be different pre- and post-2014..

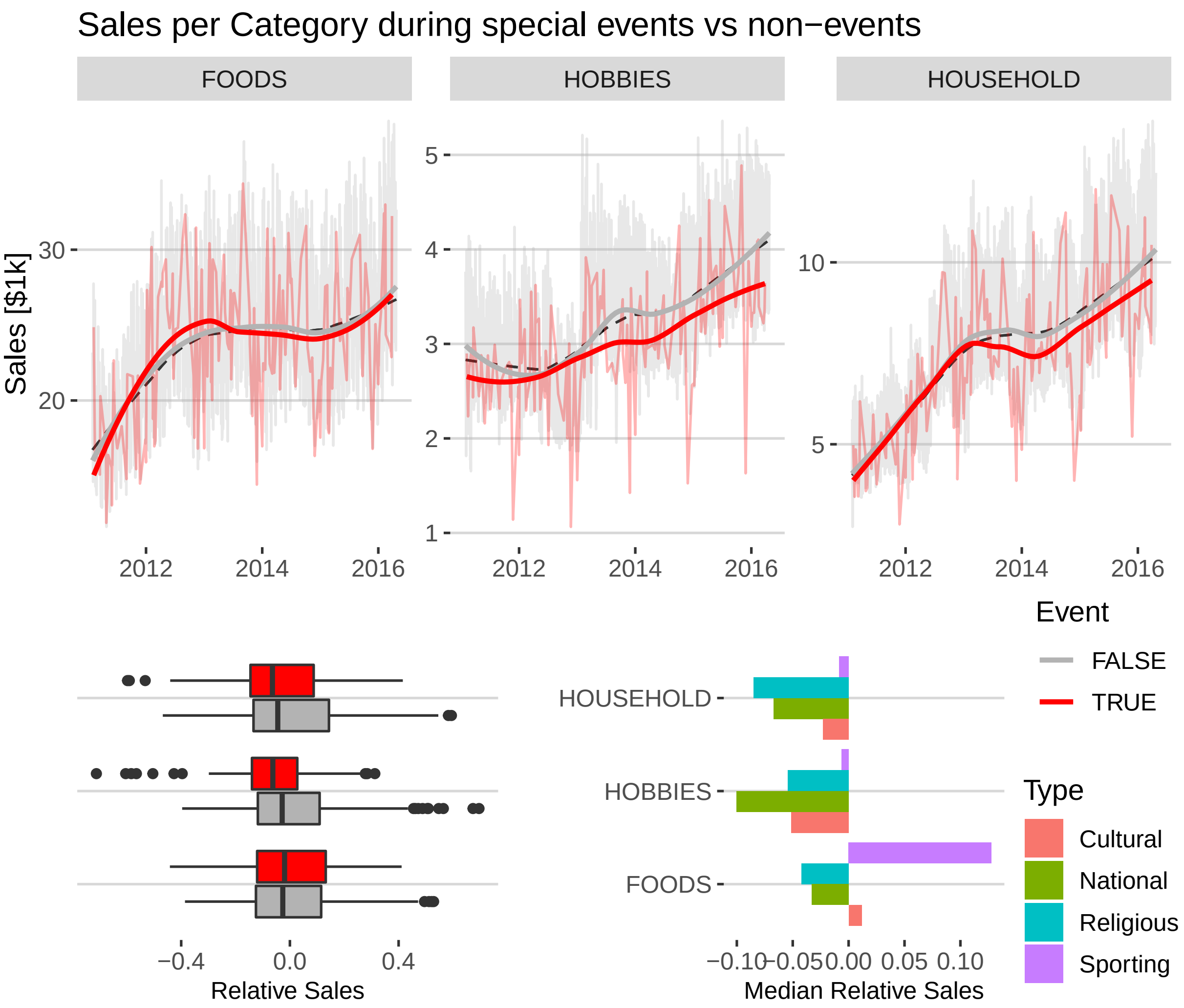

Figure 15: Fig. 14

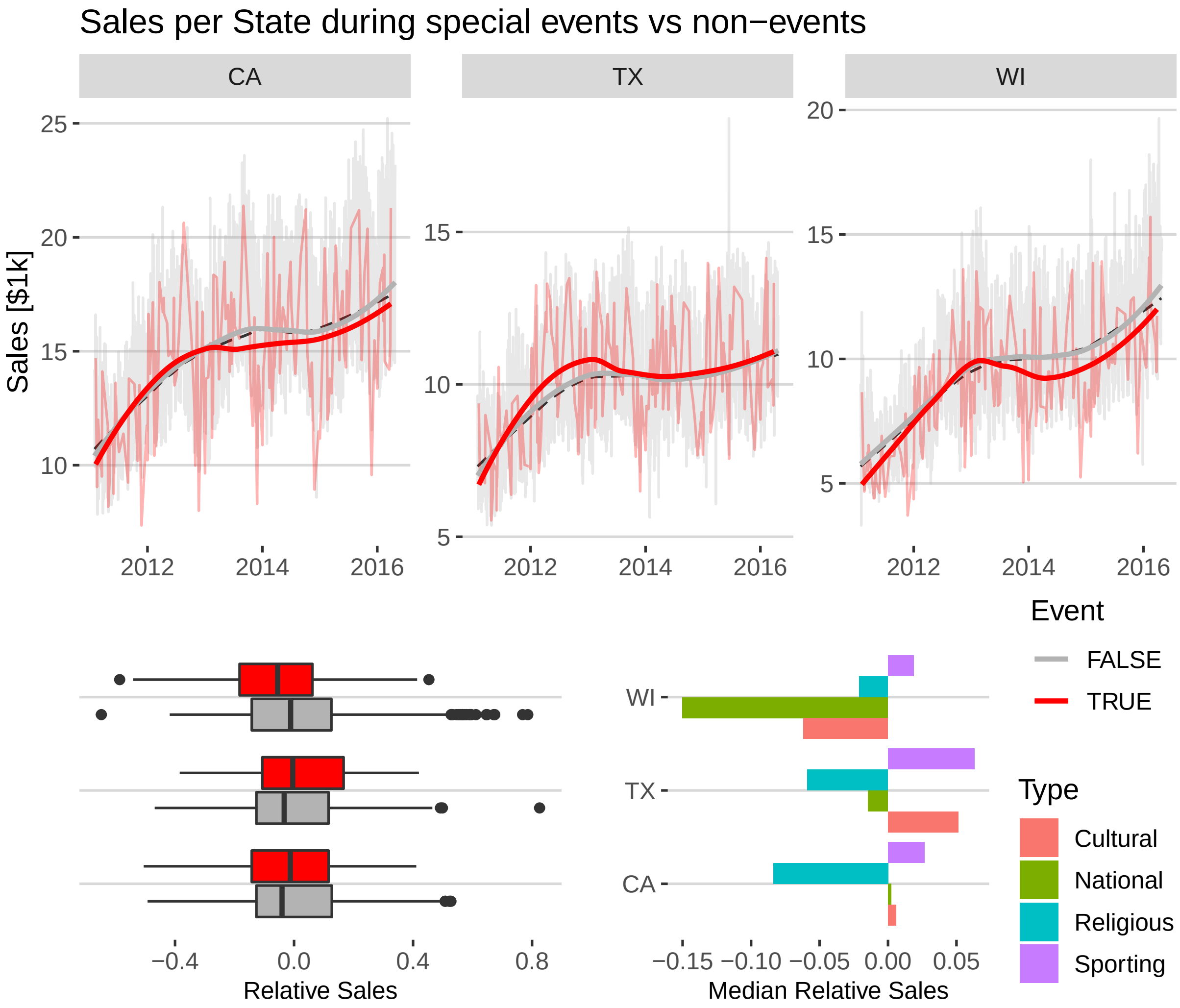

Figure 16: Fig. 15

- The view of event types looks interesting, especially for WI where "National" events have a strong negative impact on sales numbers. WI is also the only state where "Cultural" events have lower sales numbers, especially compared to TX. In contrast, "Religious" events have the smallest, but still negative impact in WI. "Sporting" events have a positive influence in each state.

Now, for the states we also have the SNAP dates from Fig. 11. In this plot we show again a smoothed fit for the SNAP days vs other days in the top panel, but then add different plots in the lower panels. The lower left plot shows the daily sales percentages. Since SNAP days make up about 1/3 of the days in each month, we sum the sales for each group (SNAP vs non-SNAP) and then divide by the total number of days.

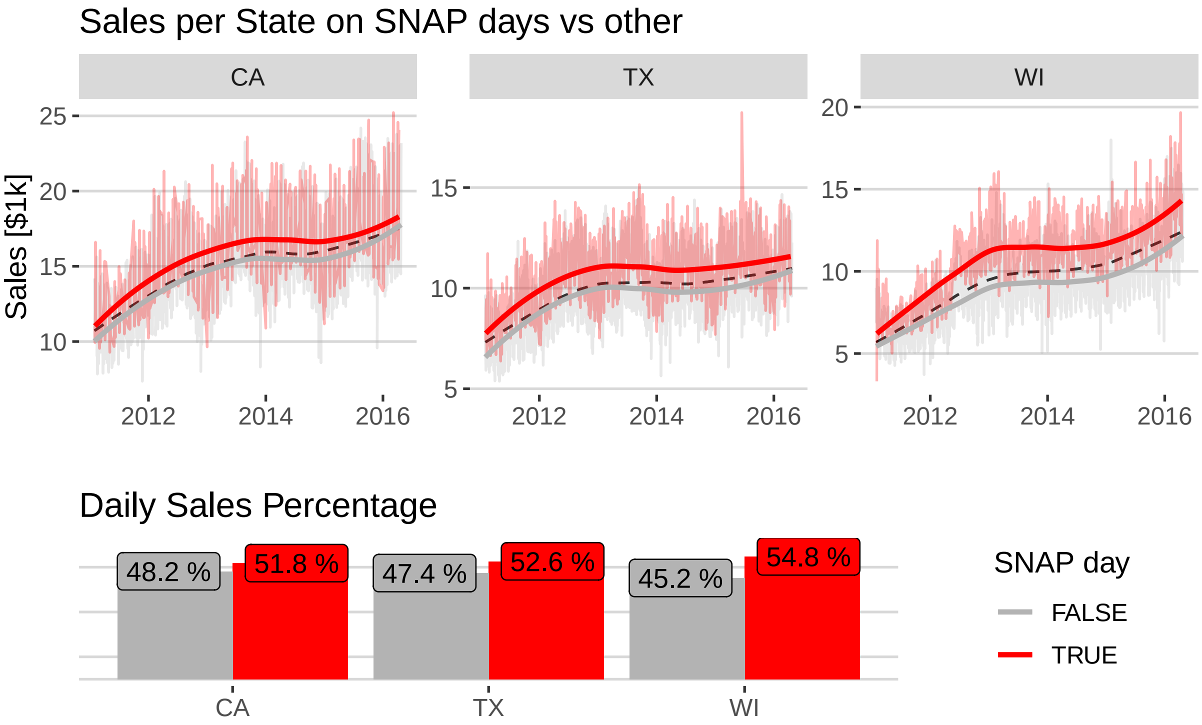

Figure 17: Fig. 16

We find:

- The SNAP days have clearly higher sales in every state. The largest difference to non-SNAP days is present for WI; this is abundantly clear from both the time series plots and the daily sales percentages. CA sales are closest for the two groups, but SNAP days are still almost 2 percentage points above 50%.
- There are some slight variations over time, primarily for WI where the two curves appear to reach there biggest difference around 2013. As with all smoothing fits, also these ones have to be taken with a grain of salt near the very edges of the data.

After looking at the last two visuals the question has to be: how does the sales `category` interact with the SNAP feature?

The "N" in SNAP stands for "Nutrition", so we would naively expect that those benefits would primarily affect purchases in our FOODS category. However, keep in mind that once someone bases their shopping patterns on SNAP days, they are very likely to purchase other items while they are there.

Let's see what the data shows. Since the SNAP days are state dependent, we have to join our data by state level (and date) first, before aggregating by category and SNAP. We first look at the daily sales percentage for SNAP vs other, and then plot a heatmap for weekday vs month. The values of the heatmap show the differences between the sum of relative sales on SNAP days minus the sum of relative sales on other days. If the numbers are positive then there are more sales on SNAP days for this weekday and month (normalised by overall volume):

We find:

- As expected, the impact is largest for the FOODs category; especially in WI. However, there are indications of slight synergy effects on other categories as well.
- The heatmap focusses on FOODS and CA (because CA has the overall largest sales numbers). We see that overall the work days Mon-Fri show stronger benefits from SNAP purchases than the weekend Sat/Sun.
- Thursdays in November stand out. I suspect that this effect is because of SNAP, but because of Thanksgiving. Thanksgiving is celebrated on the 4th Thursday in November every year. Here, this holiday most likely cuts into the "other" purchases and leads to the SNAP effect appearing artificially high.
- Which reminds us that one caveat we have to keep in mind here is that SNAP days are always during the first half of each month. Ideally, we would want to control for the possibility that the effect we see is 1st half of month vs 2nd half of month rather than SNAP vs other. Or the possibility that we see the impact of holidays like Thanksgiving instead.

4 Individual time series with explanatory variables

Now that we know the global impact of the explanatory variables, let's look at some example time series to get an idea about individual effects.

We will pick 3 random items from our interactive Fig. 2. Then we extract their sales numbers and join calendar events together with SNAP flags. I've spent some thoughts on how to display this in an efficient way, and here is my current preferred strategy: we plot the sales numbers as line charts on top of background rectangles that show the (regular) periods of SNAP days. Then we add event indicators as black points. Let me know if you have any other suggestions on how to plot this more elegantly.

To keep the focus on the SNAP periods, this plot will zoom into the period of May - Sep 2015. The three different items cover the 3 states and 3 sales categories. Following Fig. 11 we know that different states have different SNAP days; with only CA having a continuous 10-day range in our data:

We find:

- For the FOOD item, the sales patterns are consistent with more purchases during SNAP periods; as we had seen in Fig. 17. This is of course no proof of this effect. Rather we have an example of how those patterns can manifest themselves. For HOBBIES and HOUSEHOLD items there are no immediate indications that SNAP days provide a particular sales boost.
- We've seen in Fig. 15 that the impact of events is more complex. For simplicity we don't distinguish between different types of events in this plot. We see that sometimes there are sales spikes on the day of the event (e.g. FOOOS for early Jun), while other times those spikes occur prior to an event (HOBBIES late May) or thereafter (FOODS after Jul 4th). Other combinations of events and categories appear to show no particular impact either way.

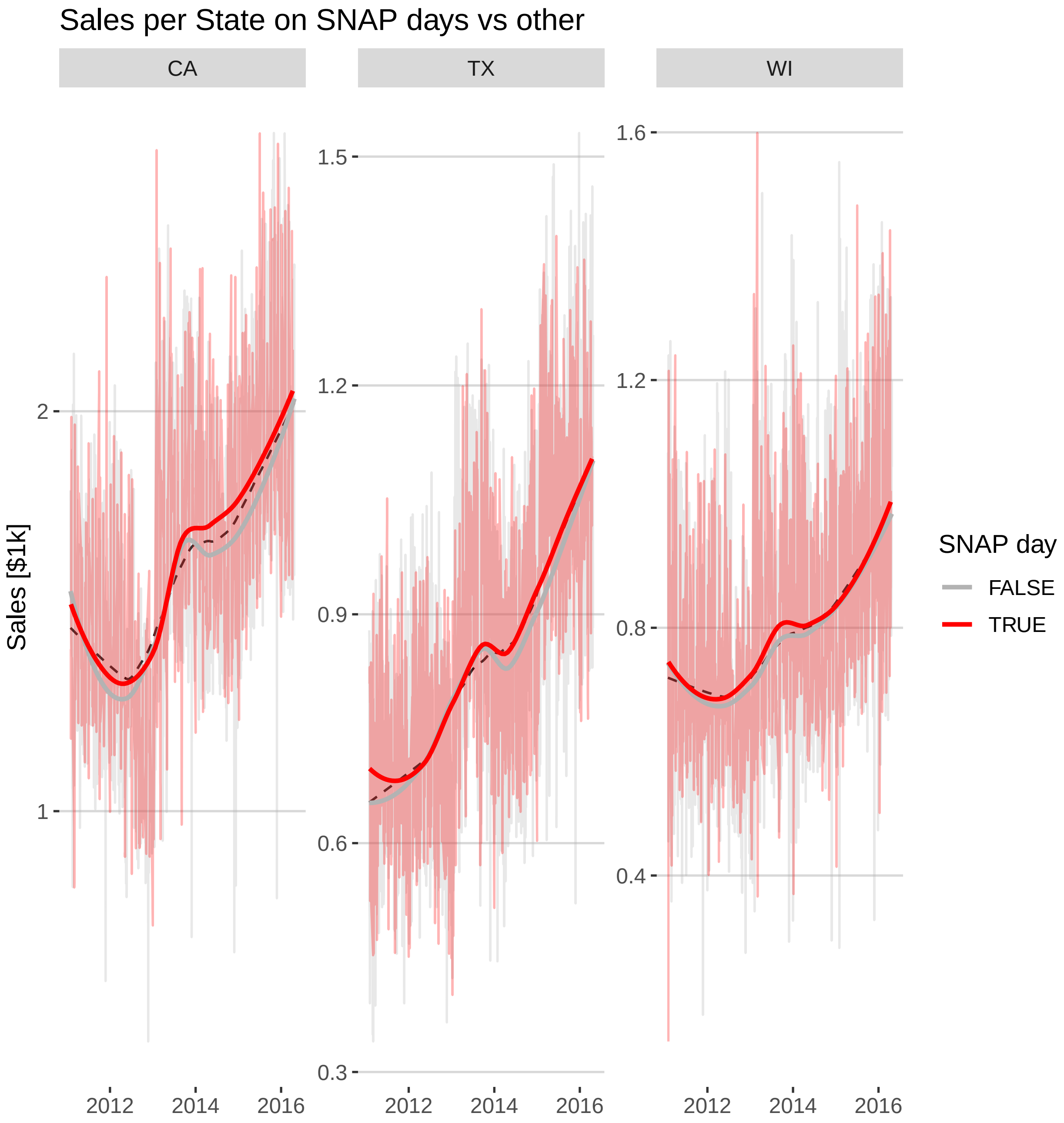

Figure 18: Fig. 17

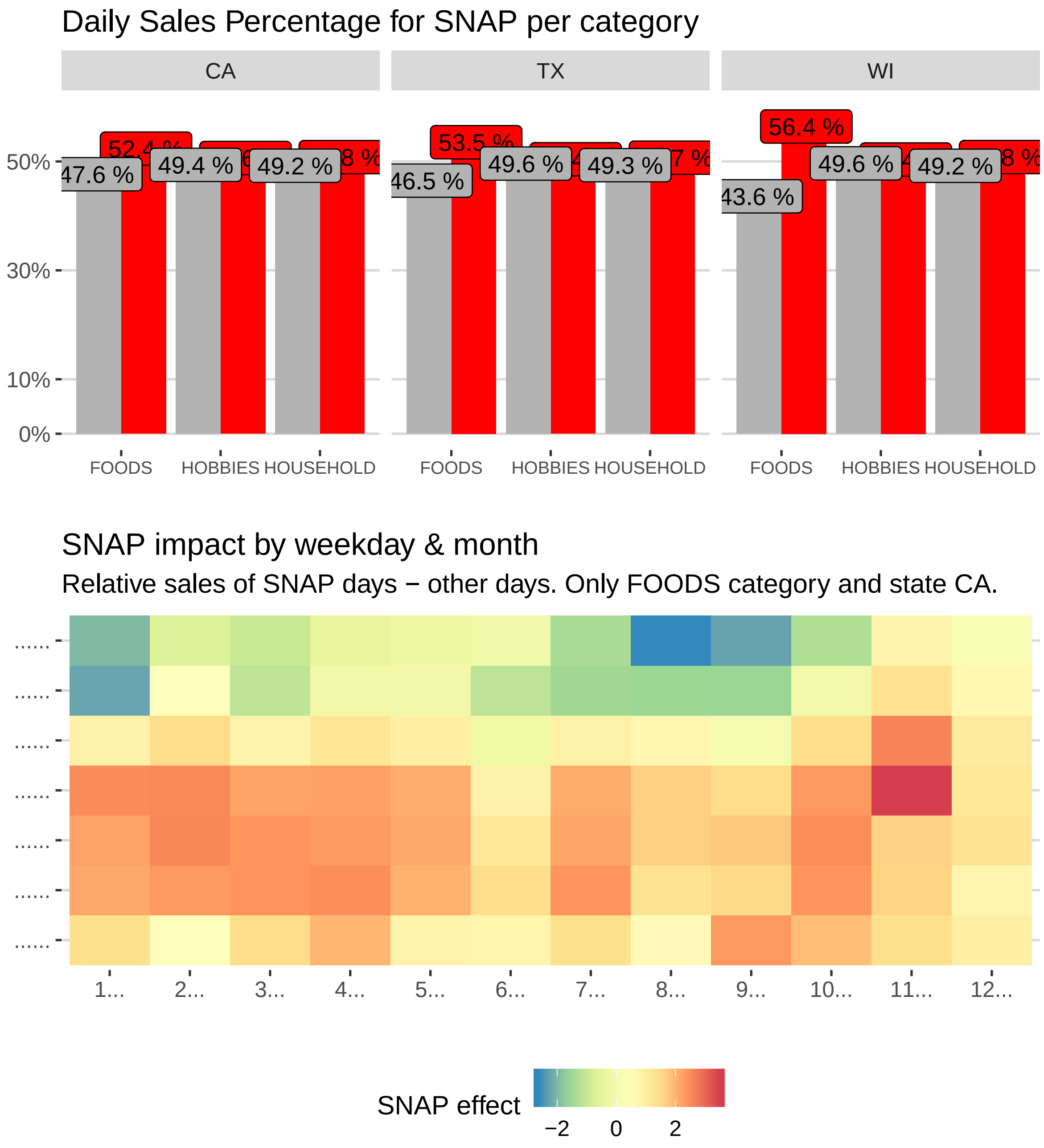

Figure 19: Fig. 17

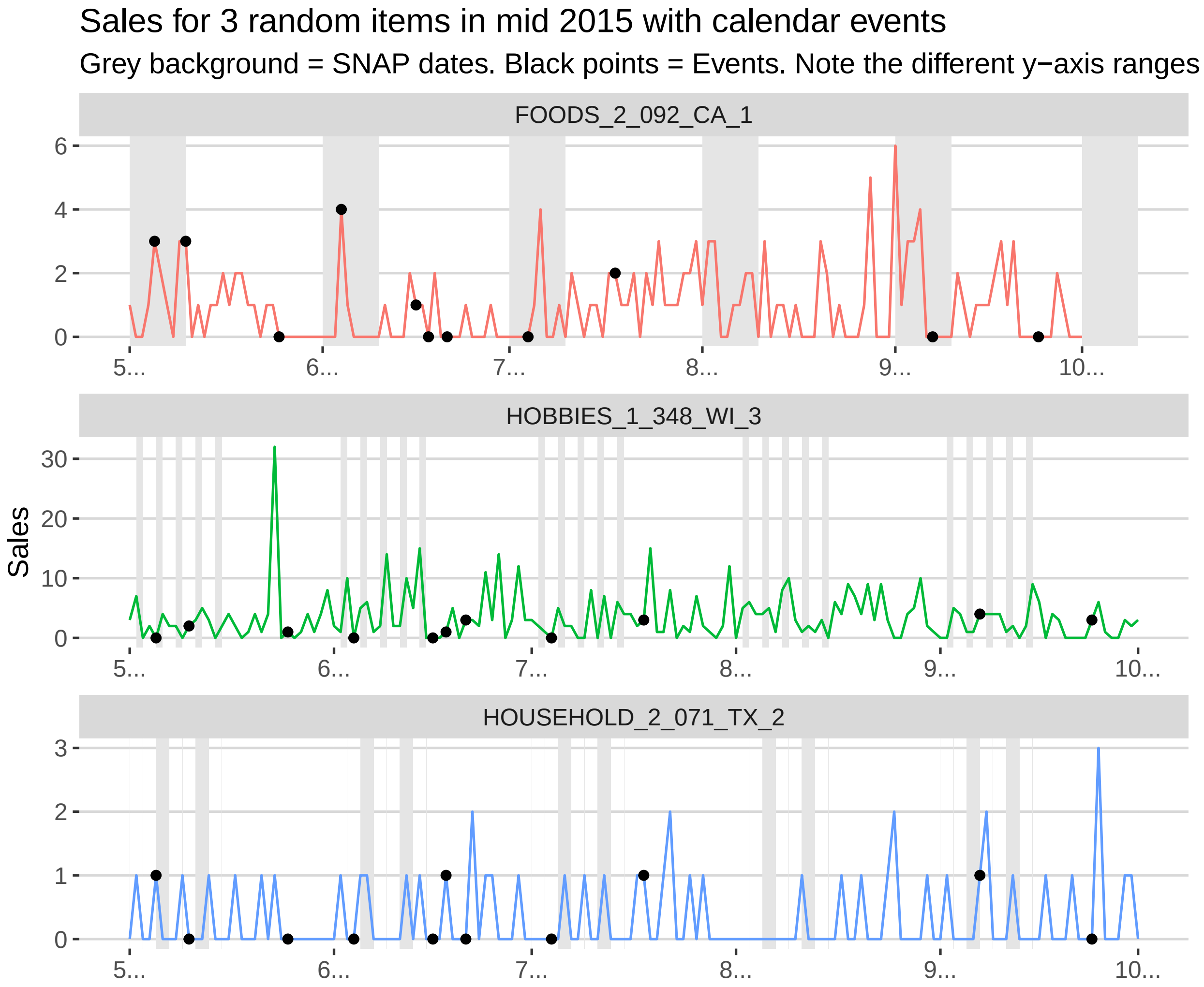

Figure 20: Fig. 18

We can look a price changes in a similar way. Here, I'm first extracting the intervals of constant prices by identifying the change points.

Then I'll join the those intervals, and their price numbers, to the sales information for our 3 example items. Since price changes happen over a longer time range, here we will look at the entire training data time range, not just focus on a few months as above. The visualisation is another experiment in which I encode the price as a background colour behind the time series. The idea is to be able to identify changes in price, whether it's an increase or a decrease, with changes in sales numbers.

In the following plots, lighter colours mean lower prices. Note the different price ranges for each item:

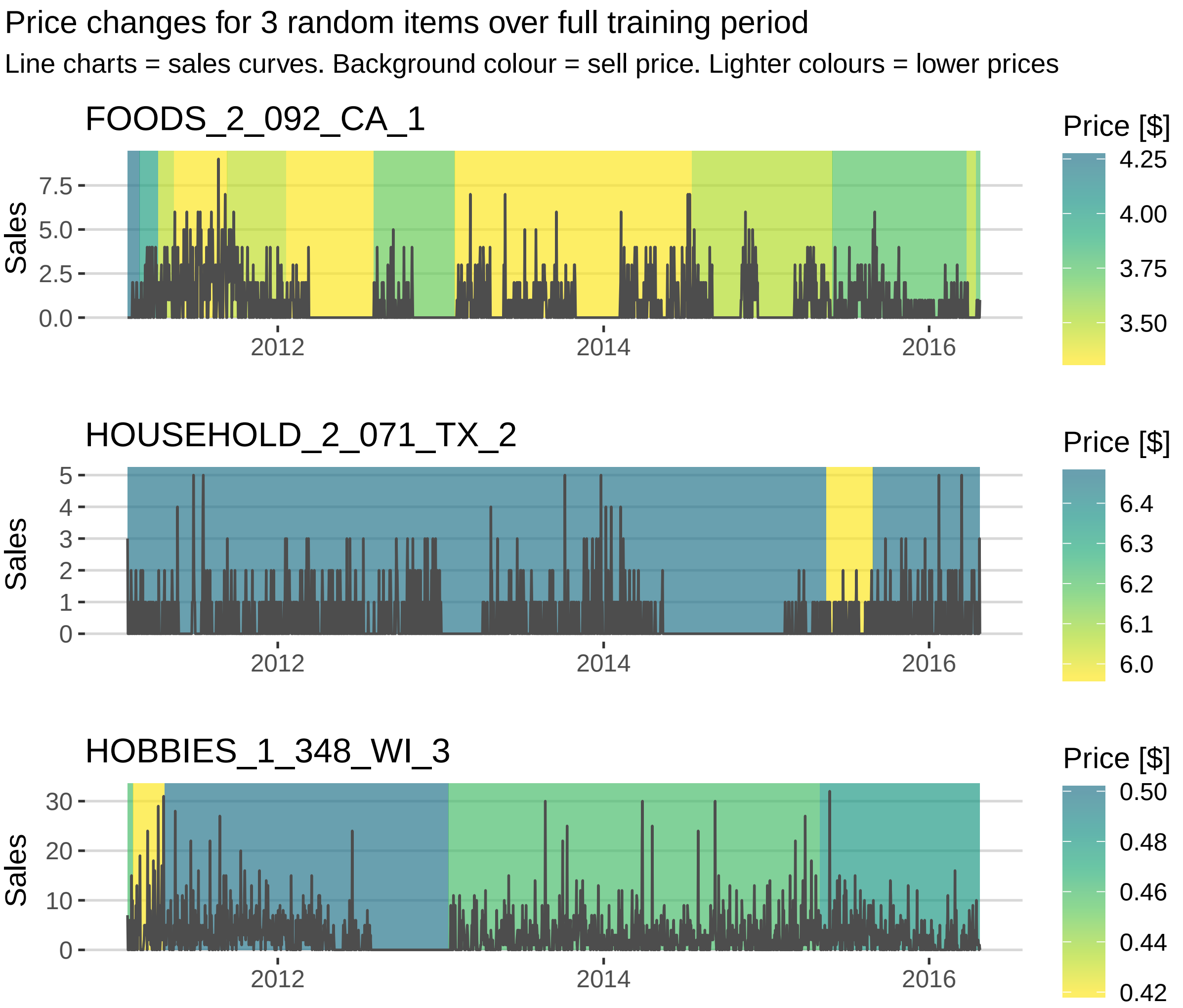

Figure 21: Fig. 19

We find:

- Some items have more frequent price changes than others. The anonymous FOODS product changes price about 10 times, whereas the HOUSEHOLD item only has one short period of lower price and then goes back to its original price.
- The main takeaway here is that in some cases price changes are correlated with demand picking up

after a noticeable gap of zero sales. Case in point: the FOODS item sales around 2013 and the HOBBIES product sales at a similar time. In both cases, the price changing during a gap coincides with the an increase in sales that lasts for a certain amount of time (rather than just 1 or 2 days). Note, though, that our data tells us nothing about whether those price changes were accompanied by increased advertising efforts, which might have more impact than the prices themselves.

- In terms of the FOODS and HOBBIES items we also see some evidence during the early parts of the time series that lower prices coincide with somewhat higher sales numbers. Keep in mind that these are simply some example time series at this point.

Alright, let's try to put everything together in another interactive plot: This one is essentially the setup of Fig. 18 for the entire training time range, with the prices from Fig. 19 overlayed as thick orange lines. The `plotly` tools allow you to zoom and pan each panel individually. In this way, you can explore each time series in different time ranges and resolutions.

Here, the vertical grey background stripes indicate SNAP dates. The black points are events, like in Fig. 18. The thick orange line shows sell prices scaled to relative values that display well in the y-axis range of the sales numbers.

Those are only 3 example items, so feel free to take this code and use it to explore other time series that you are interested in or that might be causing issues in your models. I'd be curious to know of any specific examples for which you find odd or unexpected patterns.

5 Summary statistics

Let's make a big jump from the detailed view of individual time series to a collection of overview statistics. Here we attempt to parametrise a sample of time series via a comprehensive set of fundamental parameters.

For the sake of keeping this analysis nimble we will only look at a random 5,000 time series. We can see that the statistics are reasonably representative by comparing the distribution of zero sales to the corresponding full distribution in Fig. 1 (Section 3.4).

We find:

- We briefly mentioned the high number of zero values at the beginning of this Notebook. It's certainly worth reminding ourselves that only a small fraction of time series have less than 25% zeros; and that that even having less than 50% zeros is not overly common.
- Once we remove the zeros, the mean number of daily sales is rather low: between 1 and 2 units. However, there is a small fraction that has mean sales numbers above 5 units, or even above 10. This illustrates the overall low-count nature of our individual time series.
- The price statistics are shown in shades of green: The mean item price is chiefly between 0-10 dollars; peaking around 2 or 3 dollars. As we have seen above, prices depend very much on category.
- In light green we see a visualisation of the price variability in our items. Displayed here is the difference between maximum and minimum price, divided by the mean price. Most prices don't vary much: less than 25% of their mean value. I cut off the x-axis at 100%, to focus on the interesting part, but we also see a few rare instances exceeding 200%.

Let's spend a bit more time on the zeros sales. Far from being an unimportant feature, these instances can tell us quite a lot about the properties of our time series. Here we break it down by year and look at the percentage of zeros (same as the plot above) and the year of the first non-zero instance:

We find:

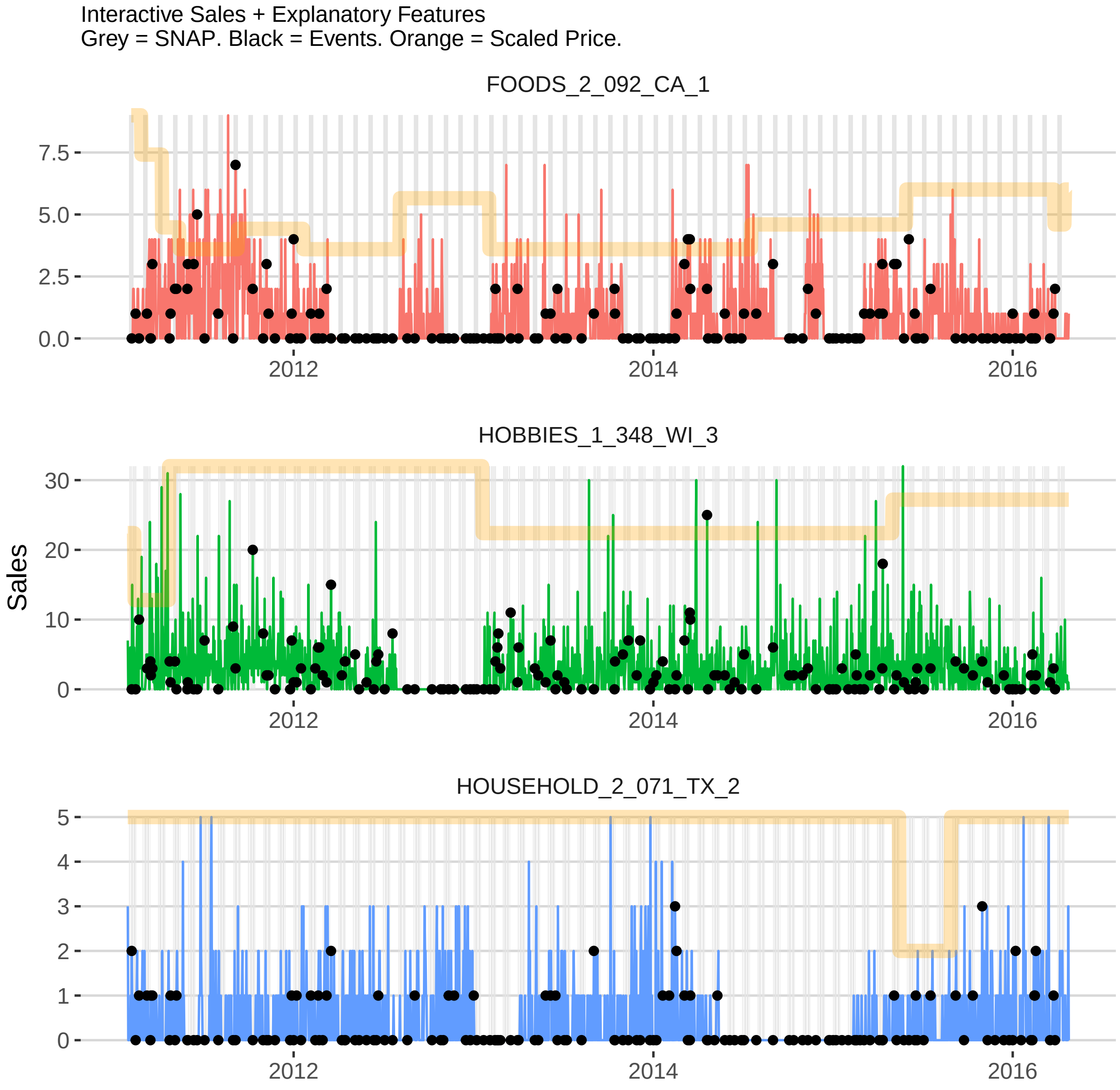

Figure 22: Fig. 20

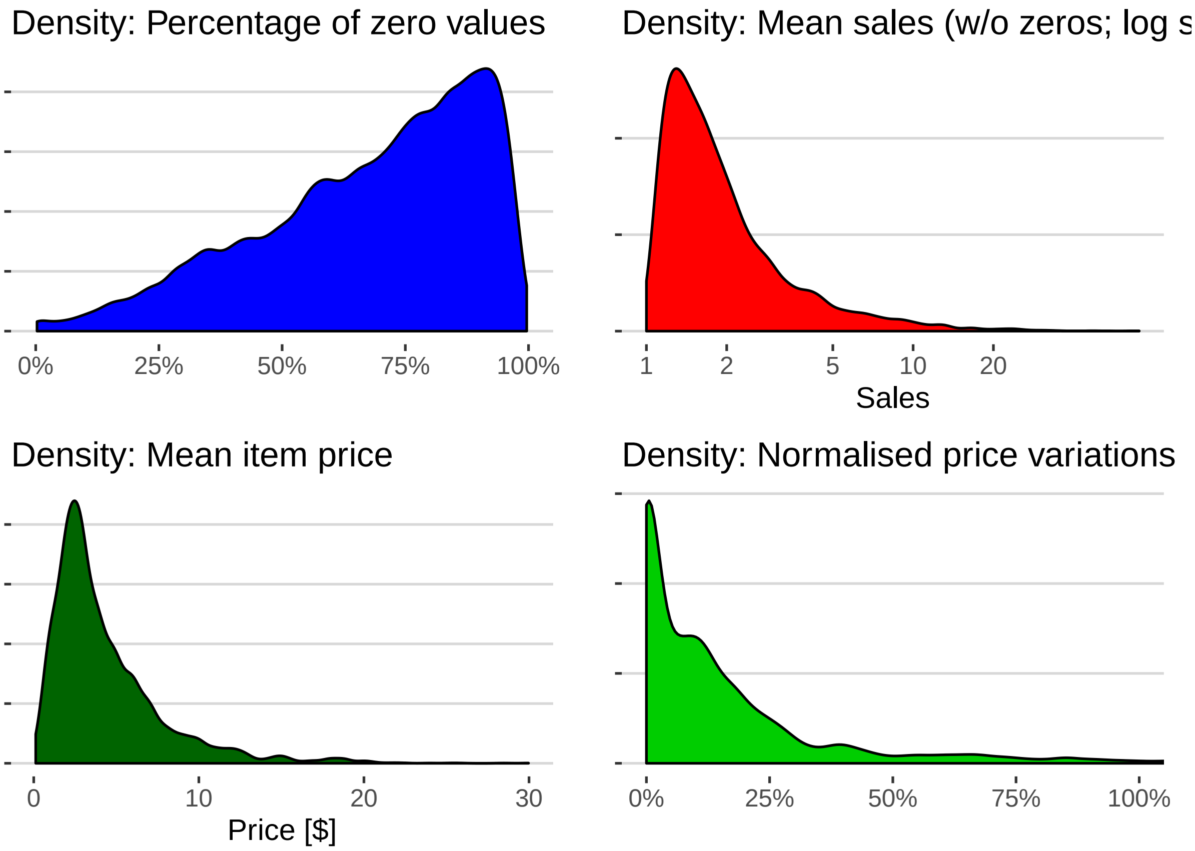

Figure 23: Fig. 21

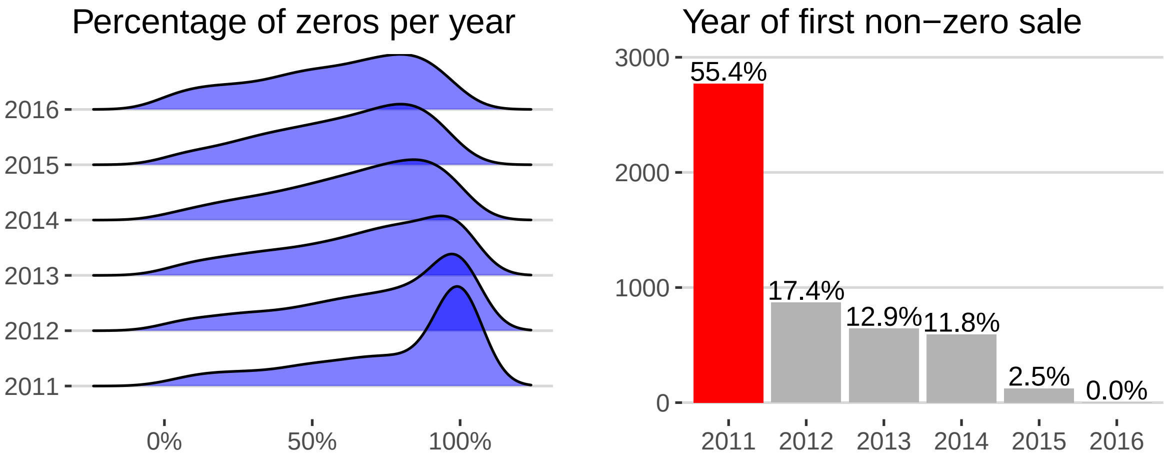

Figure 24: Fig. 22

- The percentage of zero values per year is significantly different for the first two years, 2011 and 2012, than for the remaining ones. This indicates that a notable number of time series only really start in 2013.
- This impression is confirmed by looking at the year of the first non-zero entry: while 2011 has by the single largest chunk of non-zero entries, it accounts for just above 50% of all of them. 2012 and 2013 together make up about 30% of all time series starting points. There's only 2 time series in our 5k sample that start in 2016.

A similar analysis can be done for the months of the year. Here our focus is slightly different: instead of looking at the starting year to determine how (in)complete a time series is, we now study which months show the lowest vs highest sales activities.

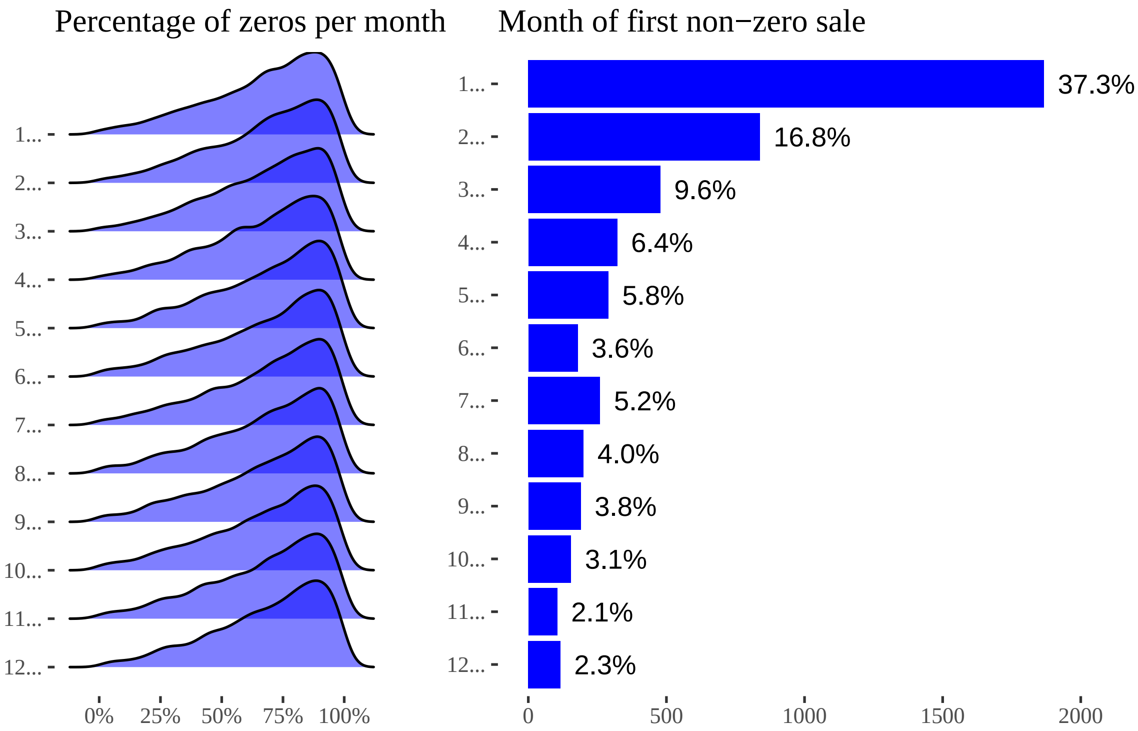

Figure 25: Fig. 23

We find:

- There's no month that really stands out in terms of zero percentage. Certainly worth a look, though.
- The right-side panel showing the month of the first non-zero entry is far more interesting. Remember that our data starts on 2011-01-29: there are only 3 days in our first year when a time series could have its first non-zero entry. And we also know that only 55% of time series have their first entry in 2011 in the first place.

This calls for a heatmap to clear up the picture:

We find:

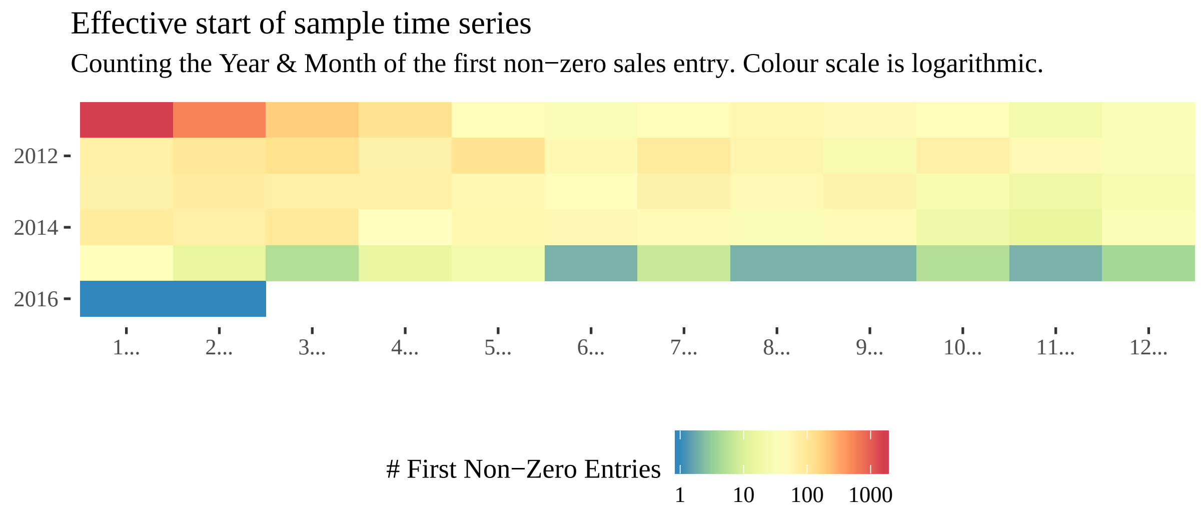

Figure 26: Fig. 24

- The single most common effective starting month for our sample data is indeed Jan 2011; despite only having 3 days in our data.
- We can also see that May and Jul have a certain edge over Jun in 2012 - 2015. This is consistent with Fig. 23 above. Since we will ultimately predict sales in Jun 2016 (see Fig. 9) this is an interesting insight to ponder.
- Similarly, Dec is elevated over Nov, but this is likely due to Holiday sales.

We can turn this question on its head, of course, and look at it from the perspective of last effective dates. A "Heads or Tails" view, if you like ;-)

Practically, this means simply looking for the maximum non-zero year/month instead of the minimum one. Everything else stays the same. Now I also add some numbers on top of the tile colours to

We find:

- The vast majority of time series extend to Apr 2016, which is reassuring. However, there are also about 250 instances (i.e. 5% of 5k) that have end dates before April. This doesn't have to mean that these items have no sales in May or June, but it does affect the forecasts.
- The total number of time series without sales in 2016 is small, but they exist.

Finally, let's have a look at price changes. In Fig. 21 we visualised the magnitude of price variations, here we'll compare their frequency and direction.

First, we will look at the number of price changes per category. This counts the number of times an item changed its price, and then plots the distribution over our three categories FOODS, HOBBIES, and HOUSEHOLD. Then, we study the direction of these price changes: we count how often the price increased and plot this number as a percentage of the total price changed. A price increase percentage below 50% indicates that there were more price drops and rises.

To display the price increase percentages we choose violin plots, which give us the global quartile measures while also preserving the shape of the distribution. Note, that since many items only changed price a few

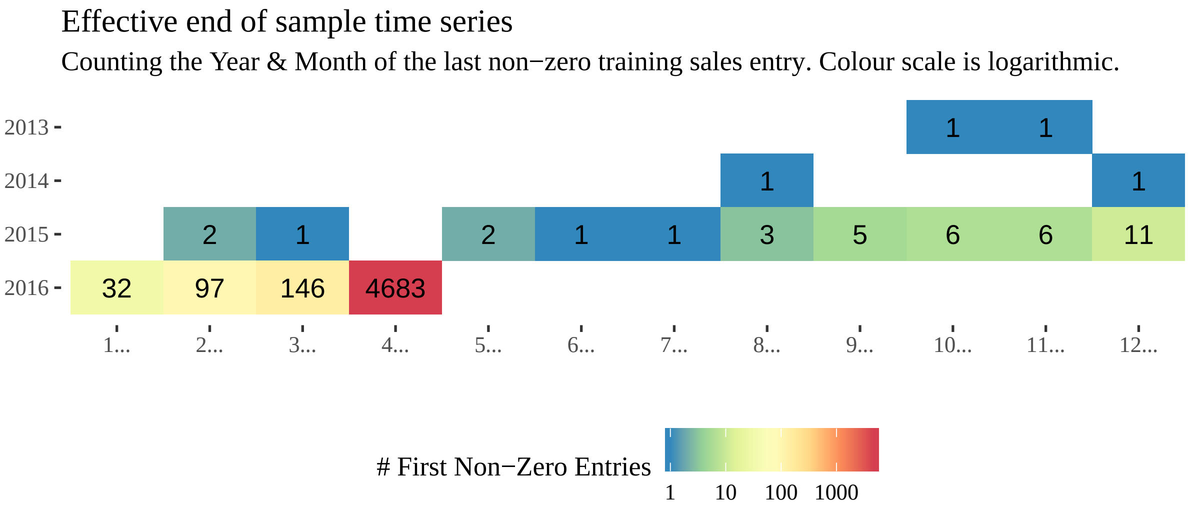

Figure 27: Fig. 25

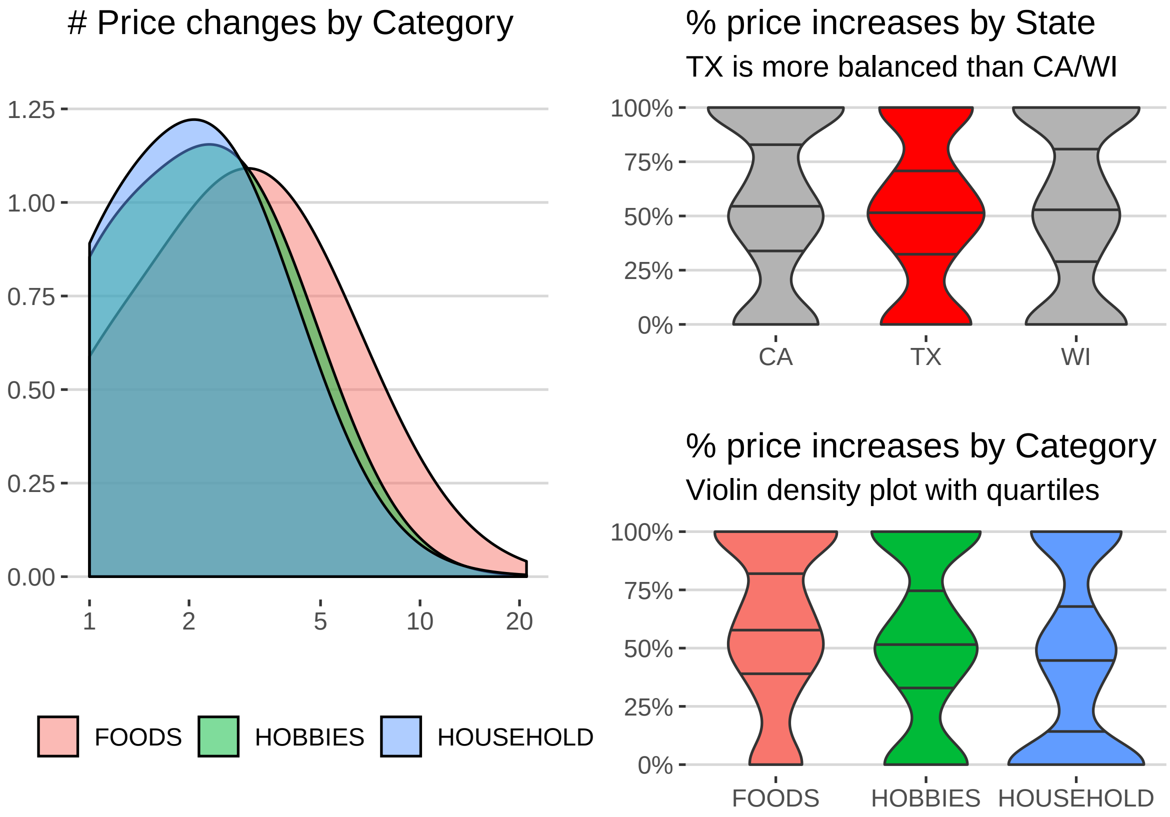

Figure 28: Fig. 26

times, the underlying data is somewhat discrete; but the violin densities do a pretty good job at visualising differences in distribution shapes. We compare price increase percentages by category and by state:

We find:

- FOODS items are seeing more frequent price changes. Overall, more than 10 price changes per item are rare.
- The FOODS category items are also more likely to increase in price over time, while HOUSEHOLD items show price drops slightly more often. The HOBBIES category is pretty balanced between items become more or less expensive.
- Between the three states, only TX shows a balanced distribution between price rises and drops. Both CA and WI are slightly more skewed towards prices increasing over time.
- Seeing that our data covers multiple years, and inflation being a thing, I would have expected prices to increase globally over time. The fact that the data shows something different is certainly notable.

Thanks !