Walmart EDA

Ella Walker

Walmart EDA

When I exploring the Walmart data I ran into several issues. I have outlined three main issues that will require adjustments. As I was conducting this EDA the inital issue that I ran into was the size of the dataset. I mutated date to be a date format because initially it was taking several minutes to load any sort of plot or figure.

1. Missing values

Table 1: Data summary

Name	big
Number of rows	421570
Number of columns	16
Column type frequency:	
character	1
Date	1
logical	1
numeric	13
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Type	0	1	1	1	0	3	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
Date	0	1	2010-02-05	2012-10-26	2011-06-17	143

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
IsHoliday	0	1	0.07	FAL: 391909, TRU: 29661

Variable type: numeric

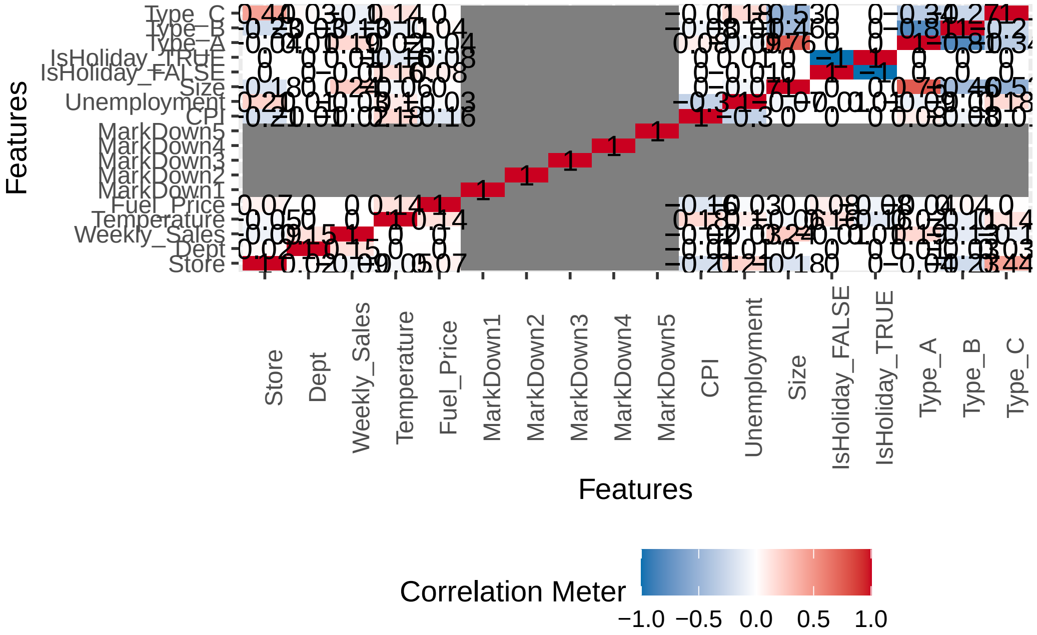
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Store	0	1.00	22.20	12.79	1.00	11.00	22.00	33.00	45.00	
Dept	0	1.00	44.26	30.49	1.00	18.00	37.00	74.00	99.00	
Weekly_Sales	0	1.00	15981.26	22711.18	-4988.94	2079.65	7612.03	20205.85	693099.36	
Temperature	0	1.00	60.09	18.45	-2.06	46.68	62.09	74.28	100.14	
Fuel_Price	0	1.00	3.36	0.46	2.47	2.93	3.45	3.74	4.47	
MarkDown1	270889	0.36	7246.42	8291.22	0.27	2240.27	5347.45	9210.90	88646.76	
MarkDown2	310322	0.26	3334.63	9475.36	-265.76	41.60	192.00	1926.94	104519.54	
MarkDown3	284479	0.33	1439.42	9623.08	-29.10	5.08	24.60	103.99	141630.61	
MarkDown4	286603	0.32	3383.17	6292.38	0.22	504.22	1481.31	3595.04	67474.85	
MarkDown5	270138	0.36	4628.98	5962.89	135.16	1878.44	3359.45	5563.80	108519.28	
CPI	0	1.00	171.20	39.16	126.06	132.02	182.32	212.42	227.23	
Unemployment	0	1.00	7.96	1.86	3.88	6.89	7.87	8.57	14.31	
Size	0	1.00	136727.92	60980.58	34875.00	93638.00	140167.00	202505.00	219622.00	

Markdown 1-5 contains promotional sales. There are thousands of missing values. However, making missing values = 0 may cause issues as the dataset states that it only has data for Markdown after November 2011 and it's not always available for all stores. This makes the missing values more complicated to deal with because it is not accurate to replace all missing values with 0.

2. Multicollinearity

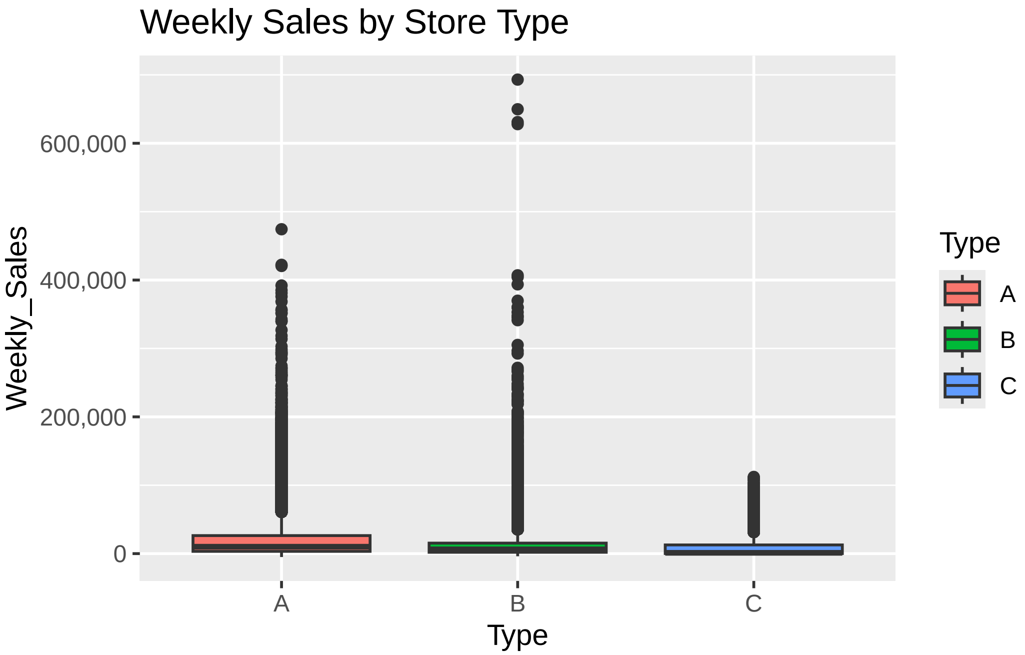
```
1 features with more than 20 categories ignored!
Date: 143 categories
```

```
Warning: Removed 150 rows containing missing values or values outside the scale range
(`geom_text()`).
```



Store type A and B are highly negatively correlated and Store A and size are high postively correlated. This does not introduce an issue in the data itself but it may cause problems if we use a linear regression model. It is logical that a certain store type would consistently be large and another type consistently small. However, we don't want to include two variables that are highly correlated with one another in a regression analysis.

3. Outliers



There are a significant amount of outliers when looking at store type and weekly sales. The data is highly skewed as we can see. However, this is also not unexpected as there are many holidays included in the dataset where we expect sales to be unusually high. If we are using a model type that has normality assumptions, like linear regression, this will be an issue. We may consider transforming the data using a log transformation.