Project documentation

▼ 0. Dataset Description

- Files
    - **train.csv** - historical data including Sales
    - **store.csv** - supplemental information about the Stores
- Data fields
    - **Store** - a unique numerical value for each store (1 - 1115)
    - **Sales** - the turnover for any given day (in $)
    - **Customers** - the number of customers on a given day
    - **Open** - indicates if the store was open (0 = closed, 1 = open)
    - **StateHoliday** - indicates a state holiday (a = public holiday, b = Easter holiday, c = Christmas, 0 = None)
    - **SchoolHoliday** - indicates a school holiday (0 = no , 1 = yes)
    - **StoreType** - differentiates between 4 different store models (a, b, c, d)
    - **Assortment** - describes an assortment level (a = basic, b = extra, c = extended)
    - **CompetitionDistance** - distance in meters to the nearest competitor store
    - **CompetitionOpenSince[Month/Year]** - gives the approximate year and month of the time the nearest competitor was opened
    - **Promo** - indicates whether a store is running a promo on that day (0 = no promotion, 1 = yes)

- **Promo2** - Promo2 is a continuing and consecutive promotion for some stores (0 = no, 1 = yes)

▼ 1. Data description and cleaning

1.1 Format columns

- rename columns from CamelCase to snake_case

1.2 Data dimensions

- verify number of lines and columns

1.3 Data types

- verify data types
  - necessary correction: date from object to datetime

1.4 Verify missing data

- variables with missing values
  - competition_distance (2642), competition_open_since_month (323348), competition_open_since_year (323348)

1.5 Fill missing values

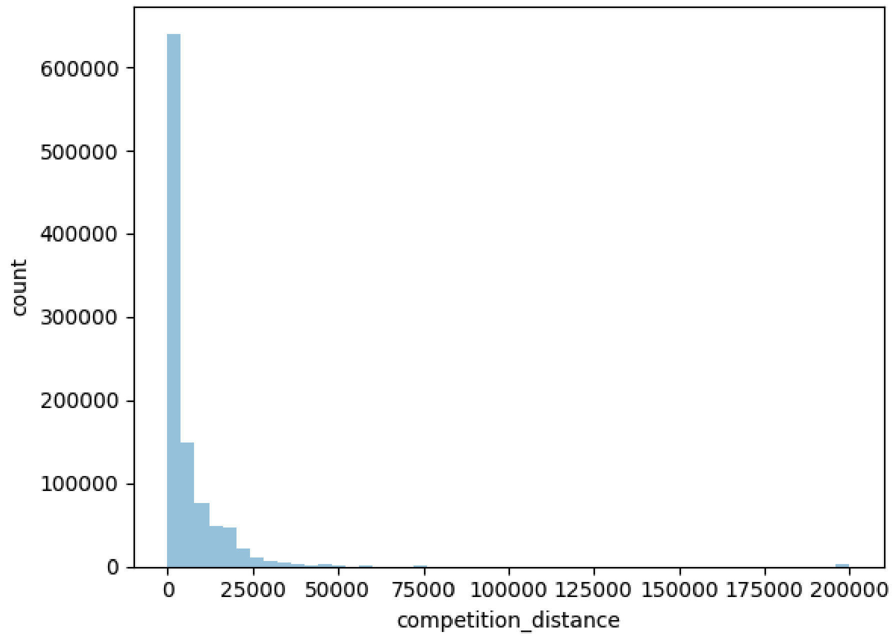
- Hypothesis:
  - competition_distance → for non-available information, it was assumed a very large distance (200000)
  - variables with date information missing were replaced with by the corresponding date sale record
    - competition_open_since_month
    - competition_open_since_year

▼ 1.6 Descriptive Statistics

- Statistical metrics
  - numerical attributes
    - highlight:

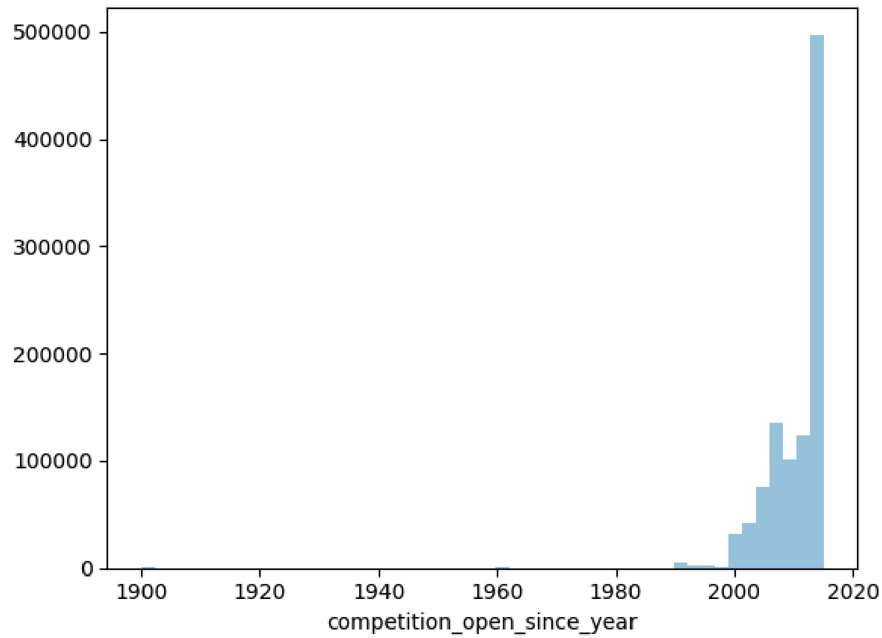
- competition_distance

  high positive skew and large kurtosis → there is a high concentration of competition near the rossmann stores

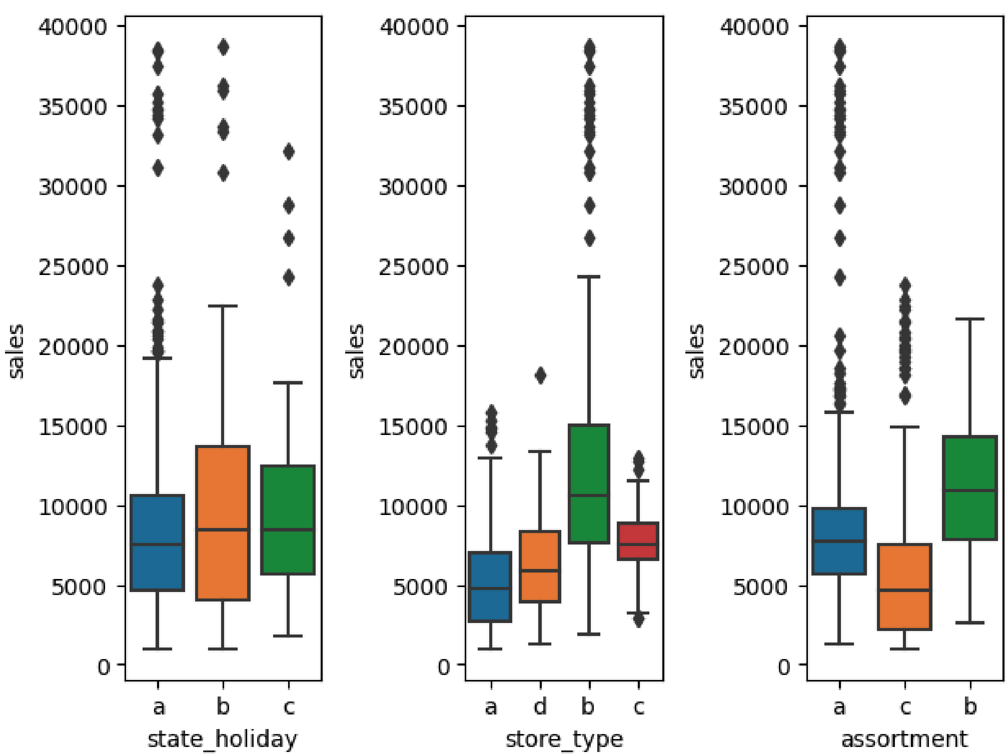


- competition_open_since_year

  high negative skew and large kurtosis → most of the competition stores opened recently



- categorical attributes

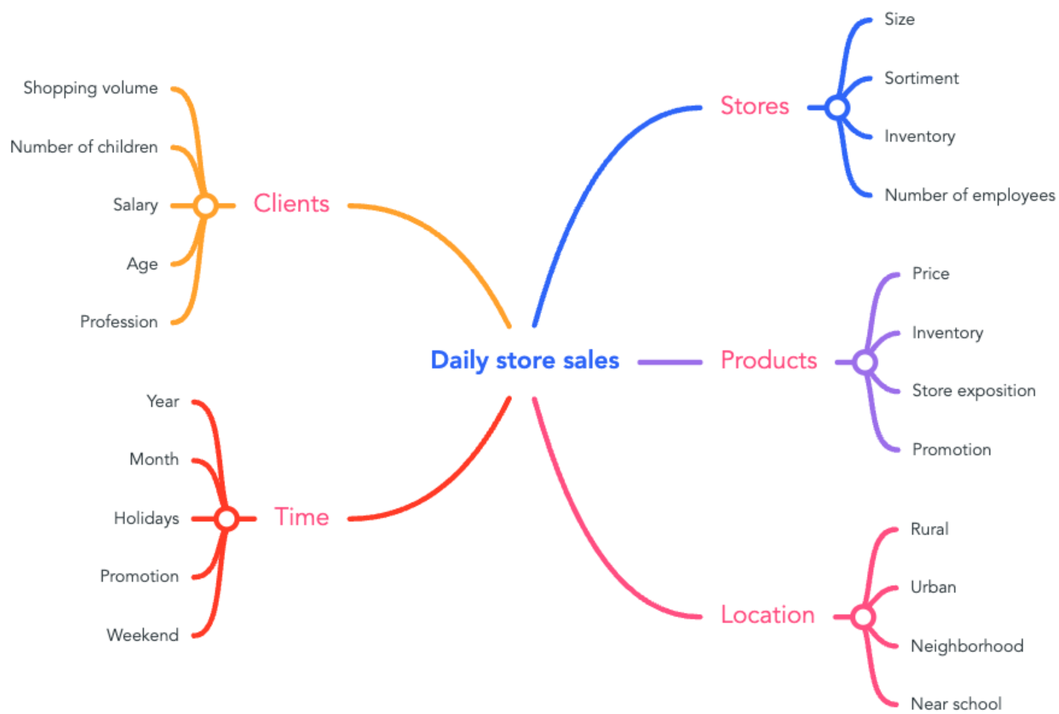


  - state_holiday

            - a = public holiday, b = Easter holiday, c = Christmas
                - in general, sales during holidays have a similar behavior
        - store_type
            - store model b has higher sales
        - **assortment**
            - the assortment level b has higher sales

▼ 2. Feature engineering

- 2.1 Business hypothesis
    - goal: raise possible questions to investigate in the explanatory analysis → therefore, evaluate if there are missing features that could be derived from the original dataset
    - hypothesis construction
        - phenomenon
            - ▼ ex: sales
        - agents
            - ▼ ex: aspects that impact sales, such as store type
        - agent's attributes



- 2.2 Hypothesis
  - based on the available dataset, the hypotheses investigate the effects of promotion, holidays, competition, store type and assortment:
    1. Stores with extended assortments sell more.
    2. Stores with closer competitors sell less.
    3. Stores with more promotion days sell more.
    4. Stores open over the Christmas holiday sell more.
    5. Stores sell more over the years.
    6. Stores sell more in the second semester.
    7. Stores sell more after the 10th of every month.
    8. Stores sell less on weekends.
    9. Stores sell less during school holidays.
- 2.3 Feature Engineering

- to test the hypothesis, the extra variables required are:
    - the day, year, and month of the sale (week of the year was also generated)
    - period of time since the competition started

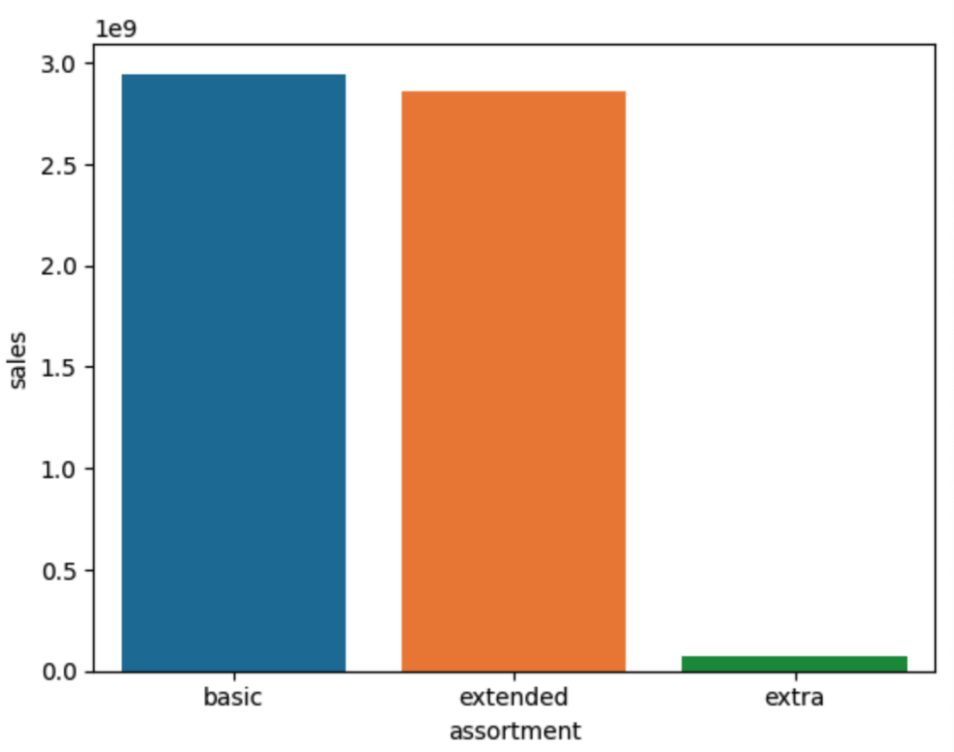
▼ 3. Data filtering

- ▼ filter the variables not available during production
    - ▼ customers: number of customers inside the store
- ▼ filter conditions that are not the interest
    - ▼ data when the store is closed
    - ▼ data when no sales were made

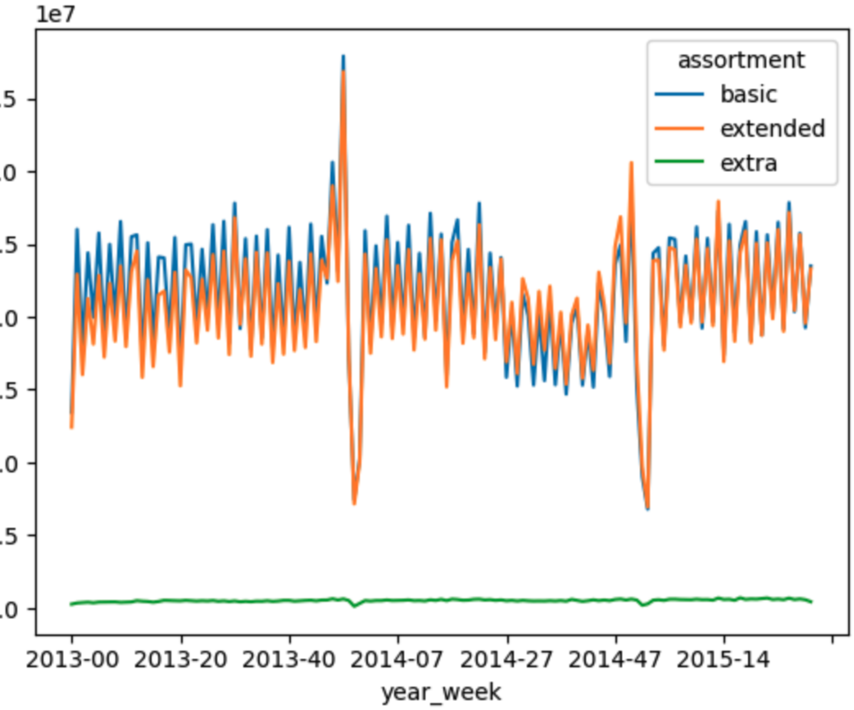
▼ 4. Exploratory analysis

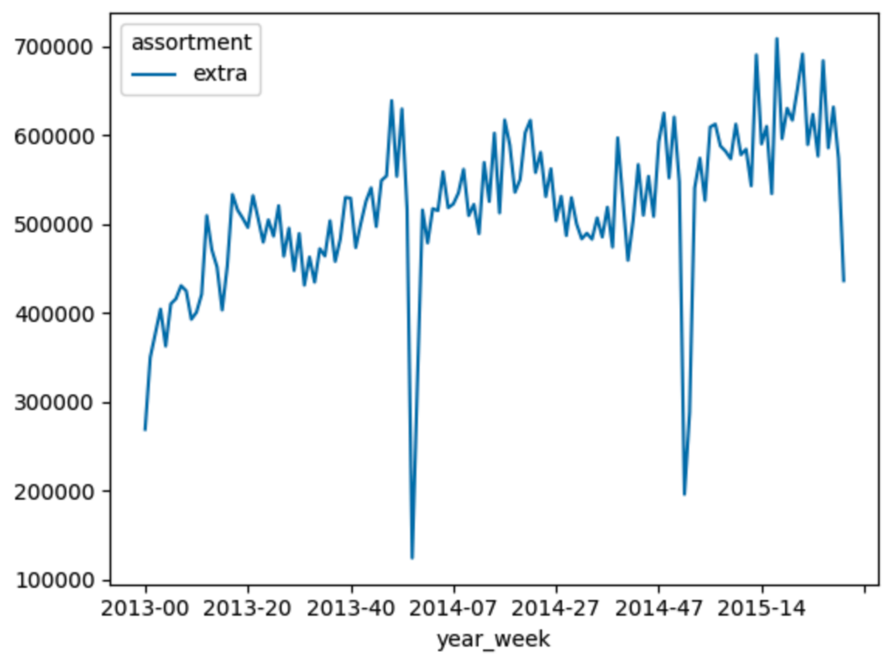
- ▼ goal:
    - business insights and hypothesis validation (item 2.1)
    - verify variables more important for the model
- ▼ 4.1 Univariate analysis
    - sales:
        - have a lognormal distribution; therefore, the logarithm is normal
    - numerical and categorical variables
        - highlights:
            - among the holidays, public holidays are the best for sales
            - store type ‘a’ makes more sales than d, followed by ‘c’ and ‘b’ at last
            - store with assortments basic and extended make more sales
- ▼ 4.2 Bivariate analysis
    - ▼ Hypothesis testing (item 2.1)
        1. Stores with extended assortments sell more.

a. False: stores with basic assortment sell more

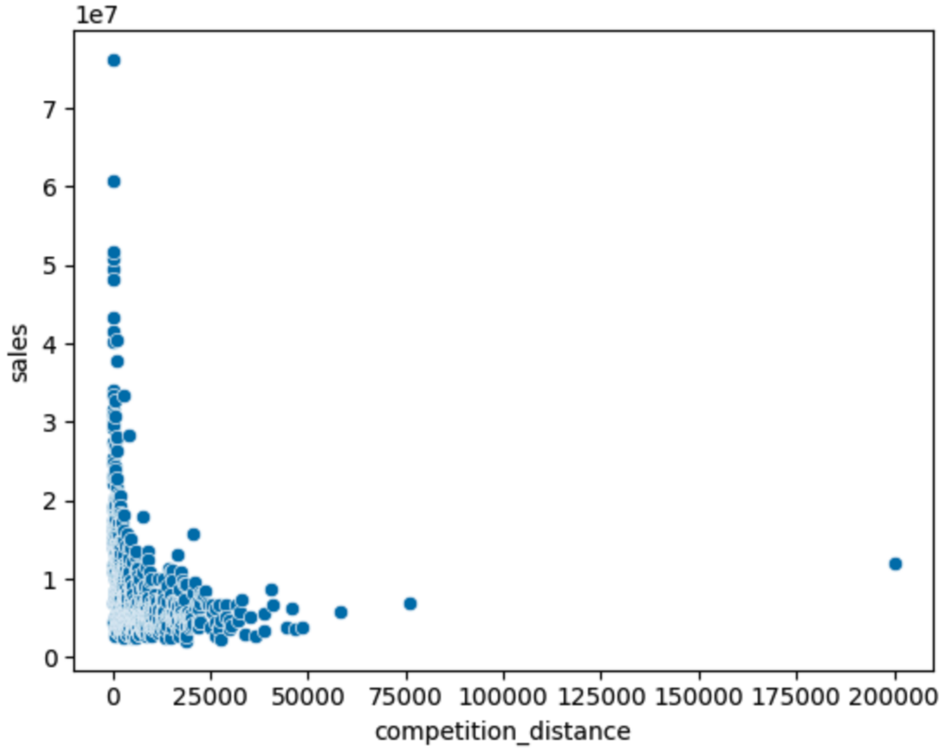


- the behavior is corroborated by weekly sales per assortment

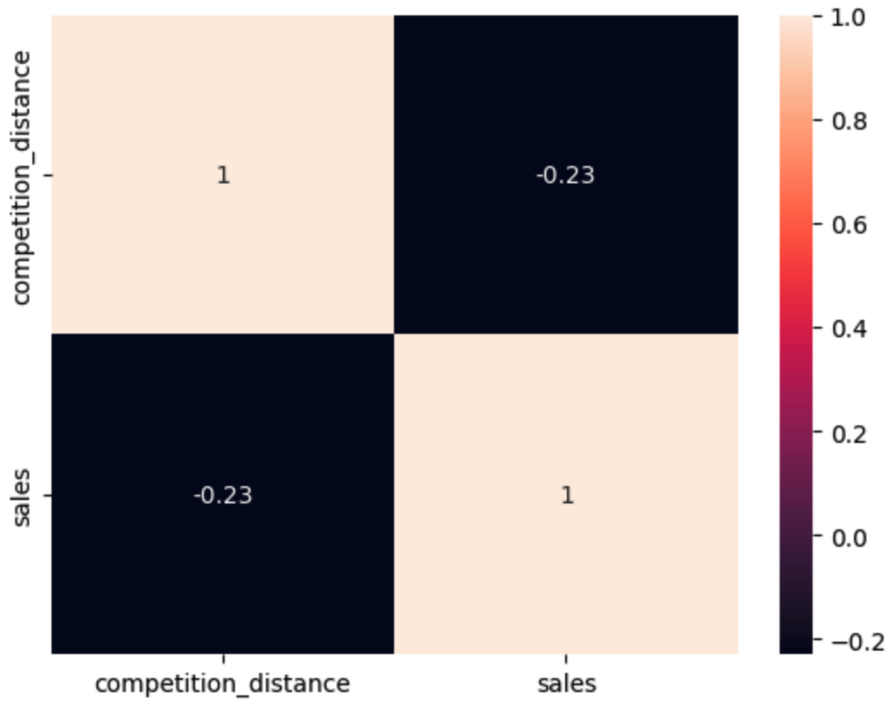




2. Stores with closer competitors sell less.
    a. False: stores with closer competition sell more



- indeed, competition distance does not explain sales (weak correlation)



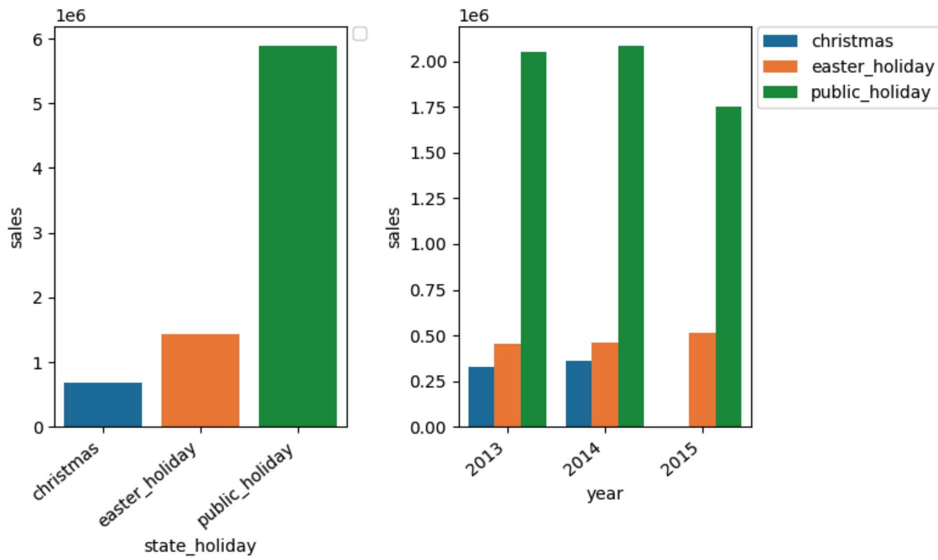
3. Stores with promotions sell more.

    a. True: during days without promotion, sales are 28% smaller, on average

    Mean sales during promo days: R$ 8228.74
    Mean sales during days without promo: R$ 5929.83

4. Stores open over the Christmas holiday sell more.

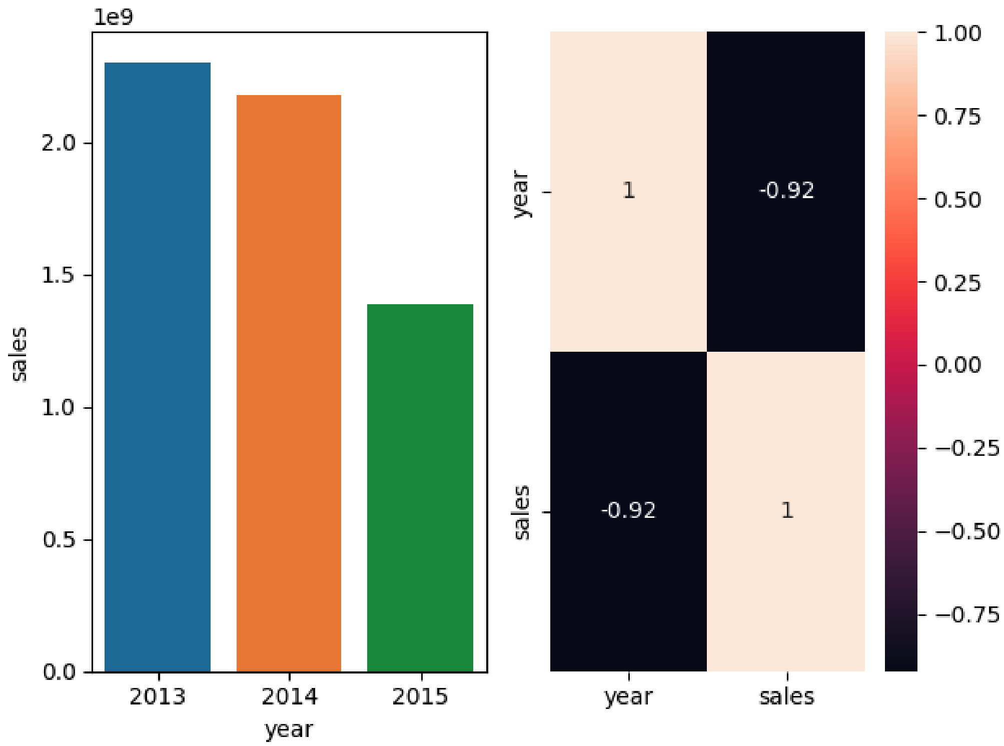
    a. False: sales are higher during public holidays during every registered year



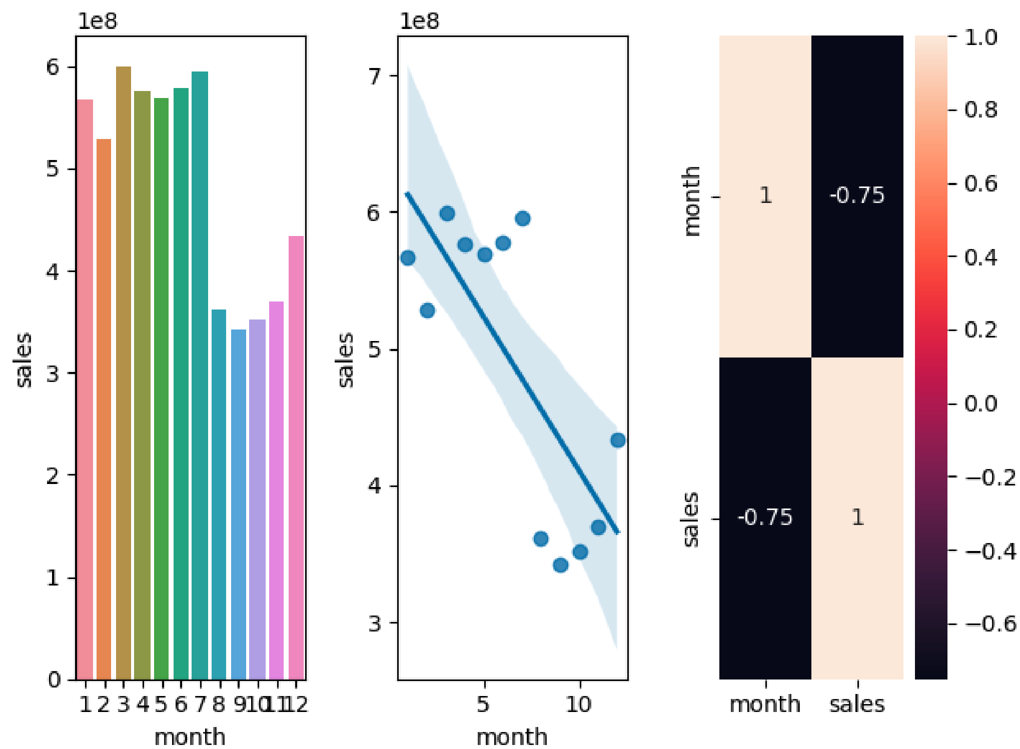
5. Stores sell more over the years.

    False: stores sell less over the years.

    Indeed, time and sales are highly negatively correlated.

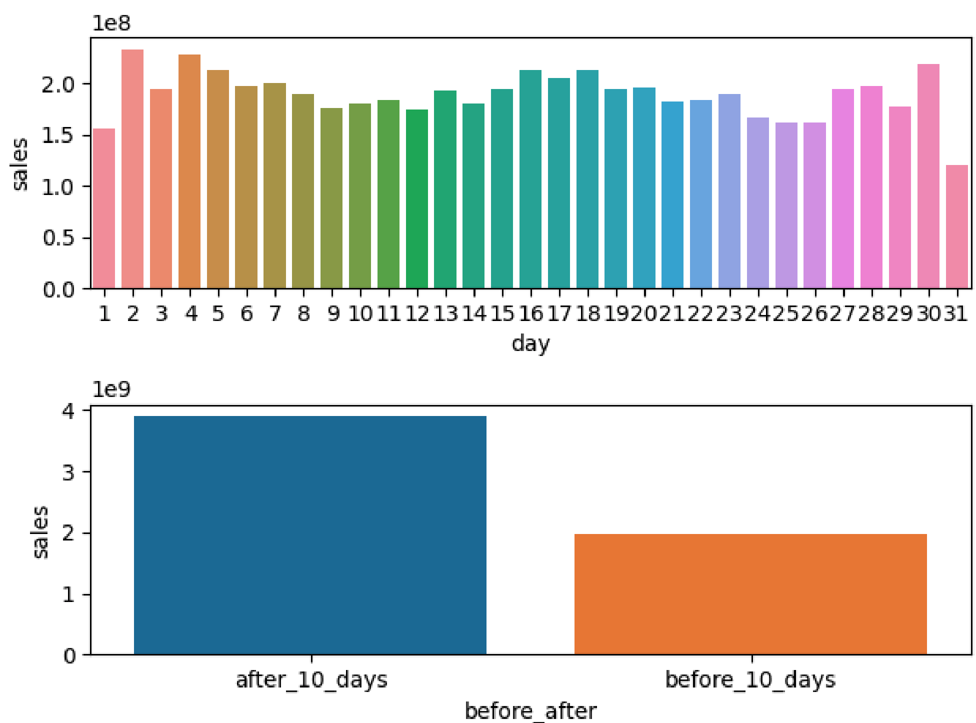


6. Stores sell more in the second semester.
    a. False: stores sell less in the second semester



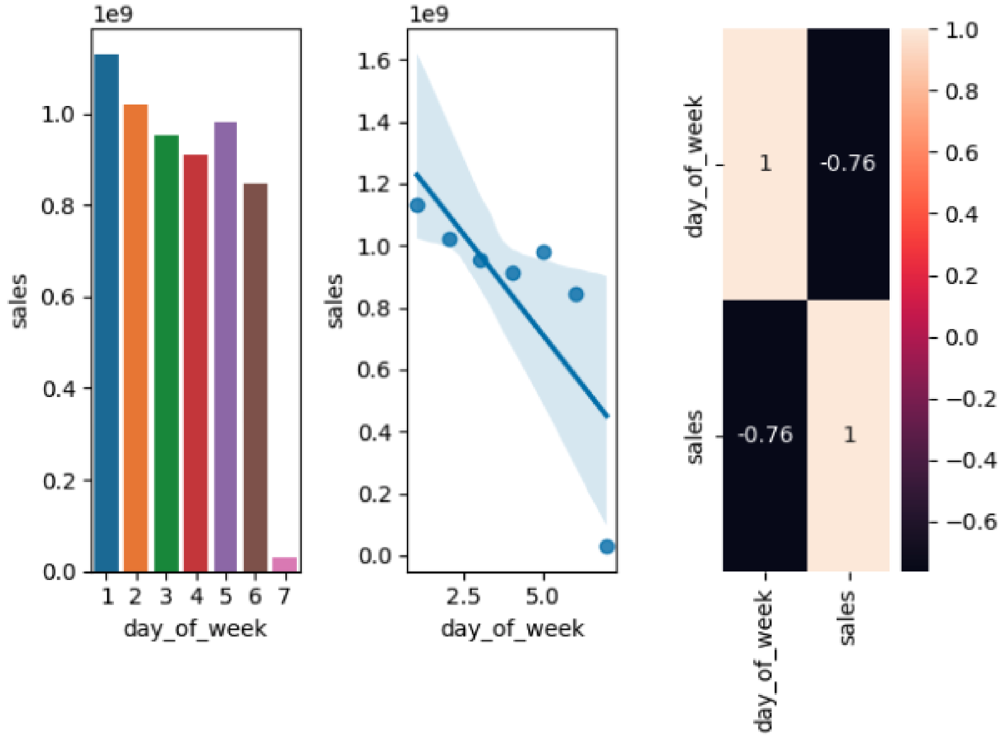
7. Stores sell more after the 10th of every month.

   True: the cumulative sales after the 10th are higher than before this day



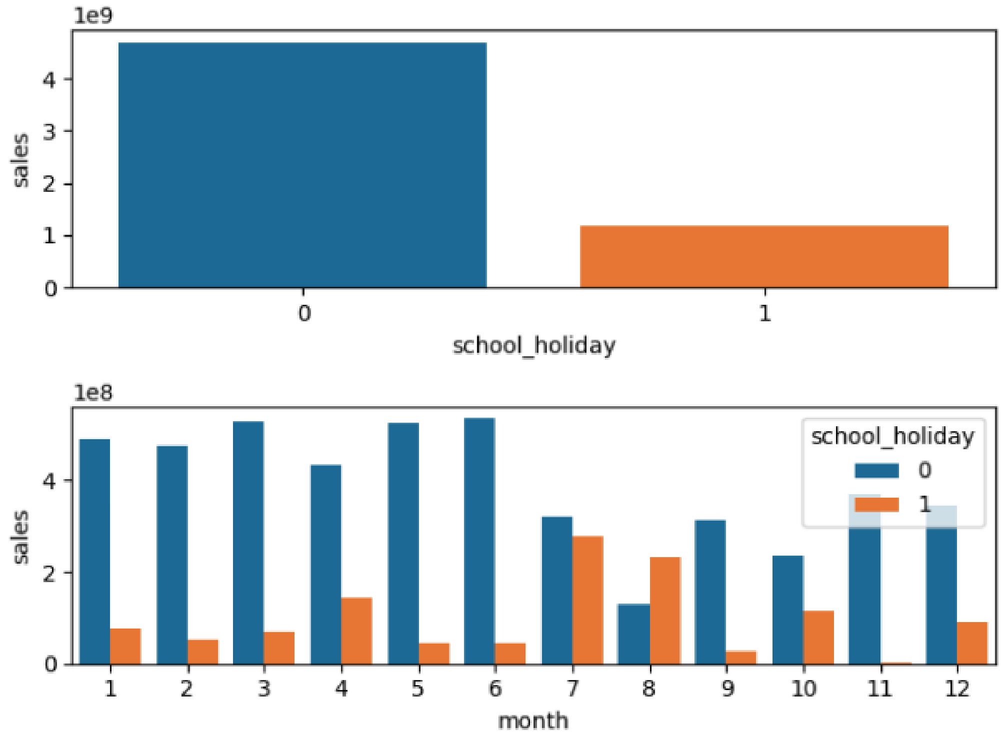
8. Stores sell less on weekends.

   True. There is a significant correlation between the day of the week and sales

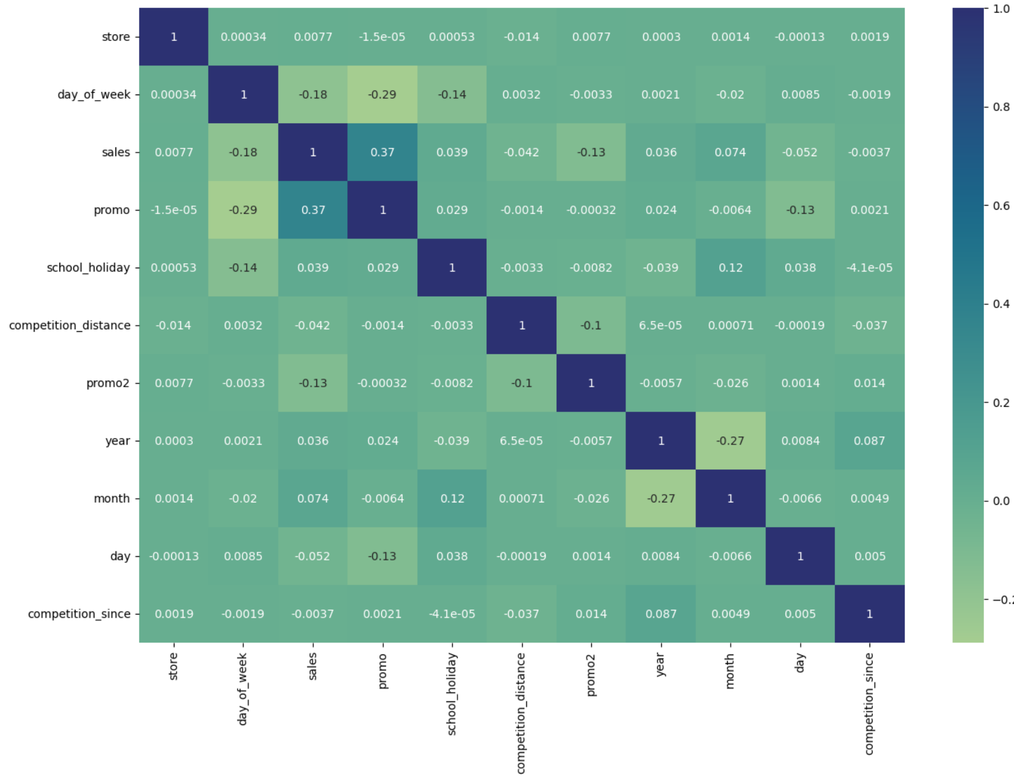


9. Stores sell less during school holidays.

   True: stores sell more on regular days (blue bars) than during school holidays (orange bars) - the exception is july - august



- 4.3 Multivariate analysis
  - Numerical attributes:
    - most correlations are weak between variables and sales are weak



- Categorical attributes
    - Cramer's V is calculated to evaluate the correlation of categorical variables
    - The results indicate a medium correlation between store type and assortment

▼ 5. Data preparation

- goal:
  - adjust variables to similar ranges for model application (otherwise the adjustments might be biased)
  - transform the var. resposta com variação mais próxima possível de normal
  - represent the distribution of cyclical variables (e.g. months): transformation
  - transform categorical variables into numerical values: encoding

▼ 5.1 Normalization: not applied, as none of the variables has a gaussian distribution (see item 4.1.2)

▼ 5.2 Rescaling

▼ applied to the variables with non-gaussian distribution (as verified in item 4.1.2): competition_distance, competition_since, and year

▼ methods:

- ▼ Robust scaler for competition_distance and competition_since
  - ▼ selected to avoid over sensibility to outliers
- ▼ Min-Max scaler for year
  - ▼ selected because it keeps the variable distribution (important because ‘year’ represents temporal evolution)
  - ▼ this method may be sensible to outliers; however, the ‘year’ variable does not have significant outliers

- ▼ 5.3 Transformation
  - ▼ encoding
    - ▼ method one hot encoding for the variable ‘state_holiday’
      - ▼ it creates new columns for the types of state_holiday, and attributes 0 and 1 as no/yes
    - ▼ method label encoding for the variable ‘store_type’
      - ▼ it attributes a random number to each store type
    - ▼ method of ordinal encoding to the variable assortment
      - ▼ it attributes a **hierarchical** number to each type of assortment
  - ▼ response variable transformation
    - ▼ sales is transformed to log, to comply with a gaussian distribution
  - ▼ nature transformation
    - ▼ variables that indicate temporal evolution were transformed as sin and cos: ‘day_of_week’, ‘month', ‘day', 'week_of_year’

▼ 6. Feature selection

- goal: select features most adequate to predict the response variable, using the Boruta algorithm
- ▼ split data into training and testing

- variables dropped:
    - the original variables must be dropped from the dataframe after the nature transformation
    - variables with temporal format:
        - competition_open_since_month, competition_open_since_year, year_week, competition_open_since (these are also redundant, as the competition_since variable compiles the information)
- test: last 6 weeks of the dataset
- train: the remaining data

▼ Boruta application

- ▼ the algorithm indicates as most relevant features the variables

▼ Feature selection

- from the Boruta and from the exploratory analysis, the following variables were then selected:
    - 'store', 'promo', 'store_type', 'assortment', 'competition_distance', 'promo2', 'competition_since', 'day_of_week_sin', 'day_of_week_cos', 'month_cos', 'month_sin', 'day_sin', 'day_cos', 'week_of_year_cos’, 'week_of_year_sin’

▼ 7. Machine learning modeling

▼ Average

- ▼ the average sale of each store is the baseline model

▼ 2 linear and 1 non-linear model

- ▼ to evaluate if sales are better predicted by linear or non-linear relationships

▼ Linear regression

▼ Linear regression regularized (Lasso)

▼ Random forest regressor

▼ Cross validation was implemented for all models

- It provides the real model performance, as the process is less influenced by the train/test split subjectivity

- Selected model:
  - Random forest model, that generated lower errors

9. Model performance

- Business implications related to model error
  - MAE (mean absolute error) ≈ R$ 665
    - Compared to the range and mean of sales in the test dataset (R$ 6995 and 40982, respectively), it is not a significant amount
  - Compared to the simple Average model (i. e. estimate the next year sales based on the mean amount sold past year), the presented model reduces the MAE from R$ 1355 to R$ 736
  - Error per store
    - The sales for some stores are more difficult to predict
      - example: store 292 has a MAPE of 63%

> This means that, if the model says that the revenue is R$100, in reality it can vary between R$ 37 and R$ 163

	store	predictions	worst_scenario	best_scenario	MAE	MAPE
922	923	399688.652290	394100.207643	405277.096938	5588.444647	1.153144
549	550	310669.399371	307471.658847	313867.139894	3197.740524	0.668210
291	292	110077.645232	106486.157906	113669.132559	3591.487327	0.627132
908	909	232321.470021	224647.263462	239995.676580	7674.206559	0.517656
273	274	168681.003557	166711.445500	170650.561613	1969.558057	0.337358

  - Total performance

- The model predicts sales for the next 6 weeks of all stores between R$ R$282,663,190.49 and R$ R$284,310,750.55

	Scenario	Values
0	predictions	R$283,486,970.52
1	worst_scenario	R$282,663,190.49
2	best_scenario	R$284,310,750.55